Vrije Universiteit Brussel

Integrated Bayesian Multi-model approach to quantify input, parameter and conceptual model structure uncertainty in groundwater modeling

Mustafa, Syed Md. Touhidul; Nossent, Jiri; Ghysels, Gert; Huysmans, Marijke

Published in:
Environmental Modelling & Software

DOI:
10.1016/j.envsoft.2020.104654

Publication date:
2020

License:
CC BY-NC-ND

Document Version:
Accepted author manuscript

Link to publication

Citation for published version (APA):
Mustafa, S. M. T., Nossent, J., Ghysels, G., & Huysmans, M. (2020). Integrated Bayesian Multi-model approach to quantify input, parameter and conceptual model structure uncertainty in groundwater modeling. *Environmental Modelling & Software*, *126*, [104654]. https://doi.org/10.1016/j.envsoft.2020.104654

Copyright
No part of this publication may be reproduced or transmitted in any form, without the prior written permission of the author(s) or other rights holders to whom publication rights have been transferred, unless permitted by a license attached to the publication (a Creative Commons license or other), or unless exceptions to copyright law apply.

Take down policy
If you believe that this document infringes your copyright or other rights, please contact openaccess@vub.be, with details of the nature of the infringement. We will investigate the claim and if justified, we will take the appropriate steps.

Download date: 17. Apr. 2024

Integrated Bayesian Multi-model Approach to Quantify Input, Parameter and Conceptual Model Structure Uncertainty in Groundwater Modeling

Syed Md. Touhidul Mustafa [1,*], Jiri Nossent [1,2], Gert Ghysels [1] and Marijke Huysmans [1]

[1]Department of Hydrology and Hydraulic Engineering, Vrije Universiteit Brussel (VUB), Brussels, Belgium

[2]Flanders Hydraulics Research, Department of Mobility and Public Works, Flemish Government, Antwerp, Belgium

*** Correspondence:**

Syed Md. Touhidul Mustafa (e-mail: syed.mustafa@vub.be)

Highlights

- Full Bayesian multi-model approach to quantify uncertainty of MODFLOW model
- Simultaneously quantifies model structure, input and parameter uncertainty
- DREAM with a novel likelihood function is combined with BMA
- Neglecting conceptual model uncertainty results in unreliable prediction
- Results in more reliable model predictions and accurate uncertainty bounds

Abstract

A flexible Integrated Bayesian Multi-model Uncertainty Estimation Framework (IBMUEF) is presented to simultaneously quantify conceptual model structure, input and parameter uncertainty of a groundwater flow model. In this fully Bayesian framework, the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm with a novel likelihood function is combined with Bayesian Model Averaging (BMA). Four alternative conceptual models, representing different geological representations of an overexploited aquifer, have been developed. The uncertainty of the input of the model is represented by multipliers. A novel likelihood function based on a new heteroscedastic error model is included to extend the applicability of the framework. The results of the study confirm that neglecting conceptual

model structure uncertainty results in unreliable prediction. Consideration of both model structure and input uncertainty are important to obtain confident parameter sets and better model predictions. This study shows that the IBMUEF provides more reliable model predictions and accurate uncertainty bounds.

Keywords: Conceptual model structure uncertainty, Bayesian approach, Input uncertainty, Bayesian model averaging, Uncertainty quantification, Groundwater flow model.

1. Introduction

The reliability of predictions of numerical groundwater flow models is strongly influenced by different sources of uncertainty. To ensure reliable predictions and decision support in sustainable water resources management, it is important to assess all different sources of uncertainty. Conceptual model structure uncertainty can be related to the complexity of a groundwater model (Elshall and Tsai, 2014), which may vary from a simple to a detailed representation of the processes and geological information of the groundwater system (Rojas et al., 2010; Mustafa et al., 2019). The geological structure is often very complex and heterogeneous and only partially known. Hence, the incomplete and biased representation of the processes, and the complex structure of a system often result in uncertainty in model predictions (Refsgaard et al., 2006; Rojas et al., 2008).

It is important to assess the different sources of uncertainty to ensure accurate predictions and reliable decision support in sustainable water resources management. The conventional treatment of uncertainty in groundwater modeling primarily focuses on parameter uncertainty, whereas uncertainties due to the model structure are often neglected (Gaganis & Smith, 2006; Rojas et al., 2008). However, many researchers have recently acknowledged that the uncertainty arising from the conceptual model structure has a significant effect on the model predictions and that parameter uncertainty does not cover the whole range of uncertainty (Bredehoeft, 2005; Højberg & Refsgaard, 2005; Mustafa et al., 2018, 2019; Neuman, 2003; Poeter & Anderson, 2005; Refsgaard et al., 2006, 2007; Rojas et al., 2008; Troldborg et al., 2007). Therefore, neglecting conceptual model structure uncertainty may result in unreliable predictions and underestimation of the total predictive uncertainty.

Most of recent studies only consider a single conceptual model structure and may fail to adequately sample the relevant space of plausible conceptual models. Single model techniques are unable to account for errors in model output resulting from structural deficiencies of a specific model as single models cannot capture all hydrogeological processes of the system (Ajami et al., 2007; Rojas et al., 2008; Mustafa et al., 2019). As a consequence, a well-calibrated model does not always accurately predict the behavior of the dynamic system (Van Straten & Keesman, 1991). Choosing a single model out of equally plausible alternative models may contribute to either type I (reject true model) or type II (fail to reject false model) model errors (Li & Tsai, 2009; Neuman, 2003).

Bredehoeft (2005) has presented different examples where the collection of new data and unforeseen elements challenged well-established conceptual models. Hence, researchers in hydrogeological science have suggested to use different alternative conceptual models for a single hydrogeological system (Højberg & Refsgaard, 2005; Mustafa et al., 2019; Nettasana et al., 2012; Refsgaard et al., 2006; Troldborg et al., 2007). Such multi-model approaches can be used to estimate a broader uncertainty band so that it is more likely to include the unknown true predicted value (Rojas et al., 2010). However, conceptual model structure uncertainty arising from the simplified representation of the hydro(geo)logic processes, geological stratification and boundary conditions, has received less attention (Refsgaard et al., 2006; Rojas et al., 2010).

A model averaging technique can be used to combine predictions of multiple models. Hydrologists have been using different model averaging techniques to obtain an average prediction and a reliable uncertainty band from a number of plausible conceptual models (Vrugt, 2016a). The predictions of multiple models are combined by using weights, which can be equal or can be determined through regression-based approaches (Yin and Tsai, 2018). Poeter & Anderson (2005) have proposed an approach in which weights are connected to model performance and the predictions of the conceptual models are combined using Akaike's weights (Akaike, 1974). However, in the multi-model predictions, this approach does not consistently include prior knowledge about parameters and conceptual models. Refsgaard et al. (2006) have proposed a method to incorporate prior knowledge of multiple model structures. In this approach, a set of conceptual models are calibrated separately and the consistency of these models was assessed using pedigree analysis. However, this method does not provide results in a quantitative way that can be used to analyse uncertainty in terms of probabilities.

On the other hand, the Bayesian Model Averaging (BMA) method (Draper, 1994; Hoeting et al., 1999) derives predictions from a set of alternative conceptual models to construct a predictive uncertainty distribution using probabilistic techniques. The weights in the BMA method are assessed based on the relative performance of each model to reproduce system behavior during the observation period. Recently, BMA has received attention of researchers in diverse fields because of its more reliable and accurate predictions than other existing model averaging methods (Li & Tsai, 2009; Rojas et al., 2008, 2010; Singh et al., 2010; Troldborg et al., 2010; Vrugt, 2016a; Ye et al., 2004, 2010).

An important challenge in implementing Bayesian Model Averaging is evaluating Bayesian model evidence (BME). There are different techniques to evaluate BME, such as analytical techniques, mathematical approximations, and numerical evaluation. The analytical solution is strongly depended on the assumptions. That is why exact and computationally efficient analytical solutions are rarely available (Schoniger et at., 2014). There are different methods of mathematical approximation, such as Laplace approximation, Kashyap Information criterion, Bayesian Information Criterion and Akaike Information Criterion. Those different mathematical information criterions may provide contradictory results in model ranking and posterior model weights (Poeter and Anderson, 2005; Singh et al., 2010; Ye et al., 2010; Schoniger et al. 2014). However, awareness about the contradictory results from different methods is very limited (Hoge et al., 2019). Although numerical methods are as prone to be biased than mathematical approximations, Schoniger et al. (2014) have concluded that bias-free numerical evaluation methods are better than mathematical approximations for model selection. Among the numerical evaluation methods, the multi-chain Markov Chain Monte Carlo (MCMC) based DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm became very popular because of its statistical robustness and numerical efficiency (Leta et al., 2015; Vrugt et al., 2008, 2016; Laloy et al., 2013;). However, applications of this algorithm for quantifying conceptual structural uncertainty of a real-world groundwater flow model also considering uncertainties from the model input and parameters are very limited.

Maximum Likelihood Bayesian Model Averaging (MLBMA), which is an approximation of BMA, has been applied recently in hydrogeology to analyse the predictive distribution of several conceptual models (Neuman, 2003; Ye et al., 2004). MLBMA depends on the calibration of alternative conceptual model parameters. However, by using this method estimated biased parameters will compensate conceptual model structure errors during calibration to obtain the best model fit (Højberg & Refsgaard, 2005; Refsgaard et al., 2006;

Troldborg et al., 2007). Refsgaard et al. (2006) have reported that the model becomes biased when calibrated models are used for simulating variables that were not included in calibration.

However, the existing Bayesian averaging approach does not quantify the uncertainty arising from the different components of the individual conceptual model and how they affect the model prediction (Tsai, 2010; Gupta et al., 2012; Tsai and Elshall, 2013). Tsai and Elshall (2013) and Chitsazan and Tsai (2015) address this issue by introducing the Hierarchical BMA (HBMA) method. In this HBMA method, the uncertainty arising from the different components of the individual conceptual model is considered using a BMA tree.

Alternative approaches to account for conceptual model structure uncertainty along with uncertainty from other sources are integrated uncertainty assessment approaches, which combine estimation of individual sources of uncertainty into an integrated modeling framework. In surface water hydrology, two distinct approaches have been developed and applied: Bayesian total error analysis (BATEA) (Kavetski et al., 2006a, 2006b; Kuczera et al., 2006) and the integrated Bayesian uncertainty estimator (IBUNE) (Ajami et al., 2007). Both methods consider model parameter, input and conceptual structural uncertainties to quantify model prediction uncertainties. However, model ranking or multi-model combinations are not considered in the BATEA framework. Hence, diagnostic model comparison is not possible in this framework. On the other hand, the IBUNE framework allows to combine multi-model predictions based on model weights obtained from a non-Bayesian optimization algorithm. As a consequence, a robust Bayesian derivation of posterior probabilities is missing. To quantify input uncertainties, the IBUNE framework uses a multiplier that is assumed to be independent and normally distributed with fixed mean and variance. Vrugt and Robinson (2007) have criticized this assumption as it is not a very appropriate way to quantify model input and conceptual structural uncertainties. Furthermore, identification of spatial and temporal variation of the input multipliers is not possible in this framework as it considers only a single multiplier. The latter might result in a biased estimation of input uncertainties and thereby result in biased predictive uncertainty. As groundwater model input data, such as recharge and abstraction rates, are usually estimated using indirect methods or specific models which are not accurate and can present errors both in space and time, the IBUNE approach is often not suitable for groundwater modeling.

In the field of groundwater hydrology, however, no systematic integrated framework has been proposed to date. Rojas et al. (2008) have applied BMA in combination with the generalized likelihood uncertainty estimation (GLUE) method (Beven, 1993; Beven & Binley, 1992) to quantify conceptual model structure uncertainty. A three-dimensional hypothetical setup with three alternative conceptualizations has been considered to demonstrate the method. However, some researchers have criticized GLUE because it is not a formal Bayesian method and may result in statistically incoherent and unreliable parameters and predictive distributions (Mantovan & Todini, 2006; Montanari, 2005; Stedinger et al., 2008). Therefore, the likelihood and threshold used for model selection and weighting in the approach of Rojas et al. (2008) has a lack of statistical basis and, as a consequence, conceptual model structure and parameters are not optimized in this method, which could result in overestimation of predictive uncertainty (Nettasana et al., 2012).

Recently, Xue & Zhang (2014) have applied multimodel ensemble Kalman filter (EnKF) in combination with the Bayesian model averaging framework to explicitly consider the model structural uncertainty. They advocated that the EnKF is computationally more efficient compared to other existing Bayesian methods. However, uncertainty arising from model input and measurement heteroscedasticity has not been explicitly considered in this framework. The performance of this multimodel EnKF framework has been tested using synthetic 2D conceptual groundwater model in idealized conditions without consideration of observational uncertainty or model bias, whereas the real-world models are often three-dimensional and more complex, and observations are not bias free (Hoge et al. 2019). Ridler et al. (2018) have also criticized this multimodel EnKF framework because of its limitation in application with bias observation. Hendricks Franssen et al. (2011) reported that the EnKF significantly outperformed with synthetic experimental data compare the real data.

Mustafa et al. (2018) presented a Bayesian approach to simultaneously quantify parameter and input uncertainty of a groundwater flow model. The performance of this approach has been evaluated using a single conceptual real-world groundwater flow model. Groundwater recharge and groundwater abstraction multipliers with a spatial and temporal character have been introduced in this study to quantify the uncertainty of the spatially distributed input data of the groundwater model along with parameter uncertainty. Nevertheless, the conceptual model structural uncertainty has not been considered in this study. As a result, the latter study is unable to account for the errors in the model output resulting from the structural deficiencies. Recently, Mustafa et al. (2019) presented a multi-model approach to quantify groundwater-level prediction uncertainty considering alternative conceptual models. In the second study, the combined effect of conceptual model structure, the climate change and groundwater abstraction scenarios on future groundwater-level prediction uncertainty has been evaluated. However, alternative conceptual models of this study have been calibrated using a local optimization method and considering only model parameter. As a result, this approach is unable to account for the uncertainty arising from the model input and parameters. Estimated biased parameters will compensate conceptual model structural errors during calibration to obtain the best model fit, as it relies on a single optimum parameter set. Moreover, the approach is missing the statistical robustness because of its deterministic modelling approach.

Very recently, Hoge et al. (2019) highlight the difference between BMA and Bayesian combined model averaging (BCMA) following Minka (2002) and Monteith et al. (2011). According to Hoge et al. (2019), BCMA means the application of equations for BMA (section 2.3) to forecast combinations of individual conceptual models instead of the application of equations for BMA to the individual conceptual model alternatives. Hoge et al. (2019) concluded that the objective of the modelling should be the main driver in selecting model averaging approaches. They also suggested to use BCMA instead of BMA if the objective of the modelling is to increase the reliability of the model prediction. The Integrated Bayesian Uncertainty Estimator (IBUNE) that has been applied in surface water hydrology by Ajami et al. (2007) has been considered as a practical application of applying BMA in similar fashion of BCMA (Hoge et al. 2019). However, as mentioned earlier, Ajami et al. (2007) allows to combine multi-model predictions based on model weights obtained from a non-Bayesian optimization algorithm. As a consequence, a robust Bayesian derivation of posterior probabilities is missing.

Hence, more research on a systematic integrated fully Bayesian framework is needed to quantify the uncertainty arising from the conceptual model structure, inputs and parameters of groundwater flow models with consideration of the heteroscedasticity of the groundwater level error. Additionally, the application of such an integrated multimodel framework on real-world cases is necessary to better understand the impacts of different sources of uncertainty on real-world model calibration and prediction problems.

The general objective of this study is therefore the development and application of an **I**ntegrated **B**ayesian **M**ulti-model **U**ncertainty **E**stimation **F**ramework (IBMUEF) to quantify input, parameter, measurement and conceptual model structure uncertainty of a fully distributed physically-based groundwater flow model to provide reliable predictions of groundwater system. In the proposed integrated fully Bayesian multi-model framework, the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm with a specific likelihood function is combined with the Bayesian Model Averaging (BMA) framework. In this new DREAM-BMA methodology, a likelihood function has been included based on the novel heteroscedastic error model for groundwater levels proposed by Mustafa et al. (2018). Like IBUNE of Ajami et al. (2007), the current study uses equations for BMA in a similar fashion as BCMA. However, unlike Ajami et al. (2007), our study allows to combine multi-model predictions based on model weights obtained from a Bayesian optimization algorithm. This is the first attempt to apply a fully Bayesian multi-model framework in real-world groundwater modeling to quantify conceptual model structure uncertainty along with uncertainties originating from model input, parameters and measurement error. In this methodology, the fully Bayesian approach proposed by Mustafa et al. (2018) has been combined with the Bayesian Combined Model Averaging (BCMA) to simultaneously quantify the uncertainty arising from the conceptual model structural, input and parameter of a fully distributed groundwater flow model. Additionally, the proposed approach is applicable for all types of residual errors i. e. both for homoscedastic and heteroscedastic errors. The IBMUEF is a flexible framework as (i) there is no limitation for the number or complexity of alternative conceptual models, (ii) users can choose the number and dimensions (spatial and temporal) of input multipliers, (iii) both quantitative and qualitative information of the system can be used in the alternative conceptual models, and (iv) it is applicable for both homoscedastic and heteroscedastic residuals errors. Moreover, the proposed approach is able to avoid compensation for conceptual model structural uncertainty arising from biased parameter estimates obtained from a model fit, as it does not rely on a single optimum parameter set.

Finally, the framework (IBMUEF) is applied in an over-exploited aquifer in the north-western Bangladesh, as it is necessary to understand the impacts of conceptual model structural uncertainties on model prediction in realistic conditions. The specific objectives of this paper are: (i) to quantify model uncertainty originating from errors in model conceptualization, (ii) to quantify individual uncertainty contributions arising from model input, parameter, and measurement and conceptual model uncertainties, (iii) to understand conceptual model structure uncertainty impacts on calibration and model prediction, (iv) to evaluate the applicability of our approach for groundwater models in realistic conditions using alternative conceptual groundwater flow models.

2. Methodology

2.1 Study area

The study area covers the six north-western districts of Bangladesh (Figure 1a). The aquifer consists mainly of medium sand, coarse sand and coarse sand with gravel, with minor fractions of clay, loamy clay, and fine sand (Figure 1c). The thickness of each stratigraphic unit moreover varies spatially. The average thickness of the top layer is 18 m and it consists of clay, clayey loam and fine sand. A 20 m thick medium sand layer is present below the top layer. The bottom part of the aquifer consists of a 35 m thick layer of coarse sand and coarse sand with gravel. Average rainfall is between 1400 mm and 1550 mm per year. However, rainfall distribution is not uniform over the year. There is almost no rainfall during the dry season (November to April), which is the major cropping season in this study area (Mustafa et al., 2017b). The area is mainly covered by irrigated agriculture of which more than 80 % is rice. Irrigated agriculture uses around 97 % of total groundwater abstraction (Shahid, 2009; Mustafa et al. 2017a). Groundwater level in this study area is continuously decreasing due to overexploitation of groundwater for irrigation (Mustafa et al., 2017a).

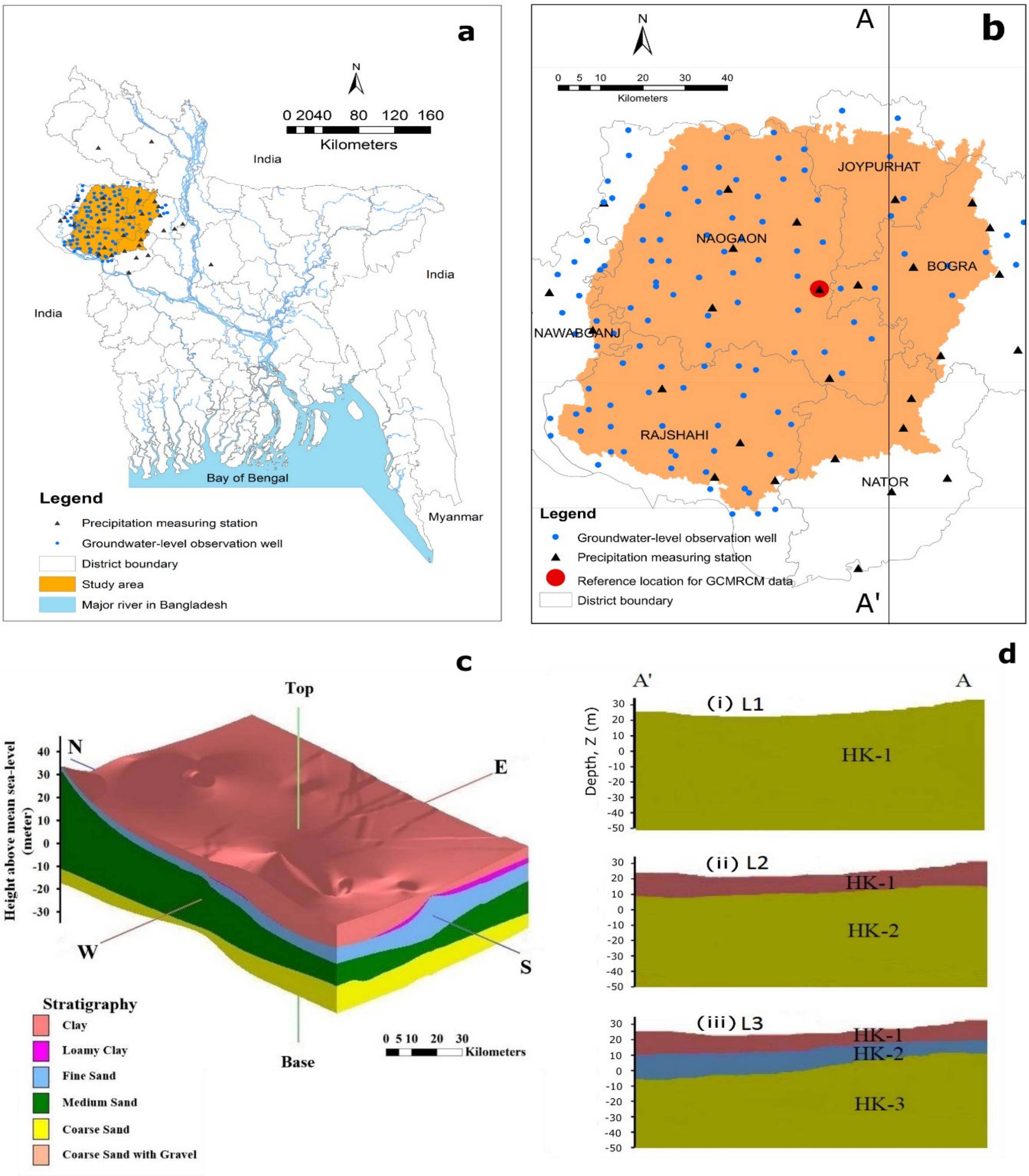


Figure 1: Description of the study area: (a) Location of the study area in the north-western part of Bangladesh; (b) study area with precipitation measurement stations (triangles) and groundwater observation wells (circles); (c) stratigraphy of the study area; (d) cross-sectional (A-A') view of different hydrogeological models: (i) one-layered model (L1), (ii) two-layered model (L2), (iii) three-layered model (L3). Taken from Mustafa et al. (2019).

2.2 Bayesian approach to quantify input and parameter uncertainty

Mustafa et al. (2018) presented a Bayesian approach to simultaneously quantify parameter and input uncertainty of a fully distributed groundwater flow model. For the details of the approach we refer the reader to Mustafa et al. (2018). A short summary of the approach is presented here. A hydrogeological model can be defined as follows:

$$O = M\ (\bar{I}, \theta, \eta) \tag{1}$$

Where $\bar{I}$ and O represent the input and output matrix of model M; θ and η are the parameters and boundary conditions of the corresponding model. To quantify input uncertainty along with parameter uncertainty, following Kavetski et al. (2002, 2006a, 2006b) a modified concept of multipliers for a fully distributed groundwater model has been introduced by Mustafa et al. (2018). The uncertainty of the input data (groundwater abstraction and recharge) is quantified using the following input error model:

$$I_{ij} = \bar{I}_{ij} * m_{ij} \tag{2}$$

Where $\bar{I}_{ij} = \{\bar{\iota}_{1,1}, \bar{\iota}_{1,2}, \bar{\iota}_{1,3}, \dots, \dots, .. \bar{\iota}_{J,N}\}$ represents the initial input for the ith month and jth location, m_{ij} is the respective input multiplier and I_{ij}represents the corresponding corrected input. m_R represents the groundwater recharge multipliers while m_A represents groundwater abstraction multipliers (Table 1). The multipliers are considered as an additional individual latent parameter and are estimated along with the model parameters.

Traditionally, residual errors in groundwater modelling are considered to be homoscedastic. However, Mustafa et al. (2018) have shown that the standard deviation of the groundwater level residual is not always constant but may increase with the deviation of groundwater level from the normal. In this study, the long-term average is considered as the normal groundwater level. A novel heteroscedastic error model for groundwater level has been proposed in this fully Bayesian approach to consider the heteroscedasticity of the groundwater level residual. The proposed heteroscedastic error model is defined as follows:

$$\sigma = A * |SH_i - \overline{OH}| + B \tag{3}$$

Where σ is standard deviation, A is a parameter representing the groundwater level uncertainty slope, B is a parameter representing the groundwater level uncertainty intercept, SH_i represents the simulated groundwater level for each time step and $\overline{OH}$ represents the observed long-term (30 years for this study) average groundwater level.

The log-likelihood function proposed by Vrugt et al. (2009a, 2013) has been adopted and modified by Mustafa et al. (2018) for spatially distributed groundwater models. The proposed novel heteroscedastic error model for groundwater level has been incorporated in this modified log-likelihood function. The new log-likelihood function is as follows:

$$\ell(\theta|\bar{I},\bar{O},\eta) = -\frac{T}{2}ln(2\pi) - \sum_{l=1}^{L}\left(\sum_{t=1}^{T} ln(\sigma_{tl})\right) - \frac{1}{2}\sum_{l=1}^{L}\left(\sum_{t=1}^{T}\left(\left(\frac{\bar{O}_{tl} - O_{tl}}{\sigma_{tl}}\right)^2\right)\right) \quad (4)$$

Where $\bar{O} = \{\bar{o}_1, \bar{o}_2, \bar{o}_3, \dots, \dots, \dots, \bar{o}_T\}$ represents the output series of observed groundwater levels in observation wells, $O = \{o_1, o_2, o_3, \dots, \dots, \dots, o_T\}$ represents the output series of simulated groundwater levels for the same observation well, $t = \{1,2,3, \dots, \dots, \dots, T\}$ represents time step, T represents the total number of time steps, $l = \{1,2,3, \dots, \dots, \dots, L\}$ represents the location of the observation wells and L represents the total number of observation wells.

This log-likelihood function has been used in this study because of (i) its numerical stability, (ii) algebraic simplicity and (iii) its applicability for both homoscedastic and heteroscedastic residual errors. To sample the posterior distribution based on the likelihood function (Equation 4), the DREAM-ZS sampler has been used. The Differential Evolution Adaptive Metropolis algorithm (DREAM) is a multi-chain Markov Chain Monte Carlo (MCMC) simulation algorithm introduced by Vrugt et al. (2008; 2009a; 2009b). The DREAM-ZS algorithm (Vrugt, 2016) has been used in this study to explicitly quantify the uncertainty arising from model input and parameters of a groundwater flow model. More details about the DREAM algorithm are explained in Vrugt et al. (2008; 2009a; 2009b) and Vrugt (2016).

In this study, we extend this approach to include conceptual model structure uncertainties and we improve the methodology by combining it with Bayesian Model Averaging (BMA).

2.3 Bayesian Model Averaging (BMA)

Bayesian Model Averaging is a probabilistic scheme for combining predictions from multiple conceptual models to provide a more realistic and reliable description of total prediction uncertainty. It is a technique that can be used to account for model structural uncertainty (Madigan et al., 1996). It is a statistical procedure that derives average predictions by weighing predictions from different models in such a way that the weighted prediction is a better representation of the observed system variables compared to any individual model of the ensemble. The BMA prediction gives higher weights to better performing models, as the agreement between the model predictions and the observations is assumed to be a measure of the model likelihood. The variance of BMA is a measure of the uncertainty of BMA prediction. The variance of BMA predictions is representing both the within-model variance and the between-model variance.

Bayesian Model Averaging (BMA) has been used to deduce more reliable predictions of groundwater levels than the predictions produced by the different individual groundwater models. Draper (1994) and Hoeting et al. (1999) present an extensive overview of BMA. Here, only a short summary of BMA is given.

Consider **M**= [M_1, M_2, M_3, … , M_K] the set of alternative conceptual models, $\boldsymbol{Y} = \{y_1, y_2, \ldots., y_n\}$ is a 1 × n observation vector of a quantity of interest, F_{jk} is the point forecast of each alternative conceptual model for $j = \{1,2, \ldots., n\}$ observations and $k = \{1,2, \ldots, K\}$ models. Now by combining the different conceptual models forecasts in a matrix **F** having dimensions of n × K, the weighted average forecast of the quantity of interest is

$$y_j = \sum_{k=1}^{K} \beta_k F_{jk} + e_j \tag{5}$$

Where $\beta = \{\beta_1, \beta_2, \ldots., \beta_K\}$ represents the weight vector of each conceptual model and e_j is noise.

As we know, model predictions are associated with uncertainty. The uncertainty can be described using a probability density function (forecast distribution) p(.). When applying BMA, assuming uniform prior distribution the posterior predictive distribution of the quantity of interest is given by

$$p(y_j|F_{jk}) = \sum_{k=1}^{K} p(y_j|F_{jk}, M_k)\, p(M_k|F_{jk}) \tag{6}$$

Where, $p(.|.)$ = conditional probability density function (PDF), $p(y_j|F_{jk}, M_k)$ = posterior predictive distribution of y_j on F_{jk} under the considered model M_k and $p(M_k|F_{jk})$ = posterior probability of the respective model M_k. This is also known as the likelihood (weight) of the corrected model M_k.

The BMA predictive mean and variance of y are conditional to the discrete ensemble of the proposed alternative conceptual models, M (Draper, 1994).

$$E[y_j|F_{jk}] = E_{\boldsymbol{M}}[E(y_j|F_{jk}, \boldsymbol{M})] = \sum_{k=1}^{K} E[y_j|F_{jk}, M_k]\, p(M_k|F_{jk}) \tag{7}$$

$$\begin{aligned} Var[y_j|F_{jk}] &= E_{\boldsymbol{M}}[Var(y_j|F_{jk}, \boldsymbol{M})] + Var_{\boldsymbol{M}}[E(y_j|F_{jk}, \boldsymbol{M})] \\ &= \textstyle\sum_{k=1}^{K} Var[y_j|F_{jk}, M_k]\, p(M_k|F_{jk}) + \sum_{k=1}^{K}(E[y_j|F_{jk}, M_k] - E[y_j|F_{jk}])^2\, p(M_k|F_{jk}) \end{aligned} \tag{8}$$

Where $E[y_j|F_{jk}, M_k]$ and $Var[y_j|F_{jk}, M_k]$ represent , respectively, the expected value and variance of y_j on F_{jk} under the considered conceptual model, M_k. Considering $E[y_j|F_{jk}, M_k] = y_k$, $Var[y_j|F_{jk}, M_k] = {\sigma_k}^2$ and $p(M_k|F_{jk}) = \beta_k$,the BMA predictive mean and variance of the quantity of interest can be developed as follows

$$E[y_j|F_{jk}] = \sum_{k=1}^{K} y_k \beta_k \tag{9}$$

$$Var[y_j|F_{jk}] = \sum_{k=1}^{K} {\sigma_k}^2 \beta_k + \sum_{k=1}^{K} \beta_k \left(y_k - \sum_{u=1}^{K} y_u\, \beta_u \right)^2 \tag{10}$$

The first term of the variance is representing the within-model variance, while the second term represents the between-model variance.

The BMA method considers the uncertainty of each model's forecast and uses it to develop a predictive distribution rather than only a weighted average. So, the BMA method provides an average forecast along with an associated forecast distribution. The forecast distribution can be used for constructing confidence intervals. This BMA forecast density enforces one significant constraint for the weights, i.e., $\beta_k \geq 0$ and $\sum_{k=1}^{K} \beta_k$=1 to avoid the development of unrealistic forecast distributions (e.g., densities can even become negative without this restriction). For successful application of the BMA method, proper estimates of the weights, and standard deviation, of the normal conditional pdfs of the ensemble members are needed. To estimate the weights and standard deviation, the log-likelihood function is used for algebraic simplicity and numerical stability,

$$\mathcal{L}(\beta_{BMA}, \sigma_{BMA}|\boldsymbol{F}, \boldsymbol{Y}) = \sum_{j=1}^{n} log\left\{\sum_{k=1}^{K} \beta_k \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\left[-\frac{1}{2}\sigma_k^{-2}\left(y_j - F_{jk}\right)^2\right]\right\} \tag{11}$$

whereβ_{BMA} is maximum likelihood Bayesian weight.

Equation (11) can only be solved iteratively. In this study, Markov Chain Monte Carlo (MCMC) simulations based on the Differential Evolution Adaptive Metropolis (DREAM) algorithm are used to calculate the log-likelihood function. The value of β_{BMA} was used as a criterion to select better performing models that have a significant contribution in model averaging.

2.4 Integrated Bayesian Multi-model Uncertainty Estimation Framework (IBMUEF)

In this framework, the fully Bayesian approach using input uncertainty multipliers based on a specific heteroscedastic error-model as explained in section 2.2 is combined with the Bayesian Model Averaging (BMA) framework explained in section 2.3. The IBMUEF framework is implemented as follows:

1. A number of alternative conceptual hydrogeological models are proposed based on the existing geological and hydrogeological information about the study area.
2. Along with parameter uncertainty, the input uncertainty of the spatially distributed input data are quantified by using groundwater recharge and groundwater abstraction multipliers (Section 2.2 and Mustafa et al., 2018).
3. A heteroscedastic error model is defined to quantify the heteroscedasticity of the groundwater level residual (Section 2.2).
4. Hydrologically reasonable prior ranges are defined for the model parameters, input multipliers and heteroscedastic error model parameters of each model (assuming a uniform prior distribution).
5. A likelihood function is defined. The likelihood function is explained in section 2.2 and Mustafa et al. (2018).
6. The posterior distributions of model parameters, input multipliers and the heteroscedastic error model parameters are obtained for each model after convergence using DREAM.
7. A pre-specified number of outputs (e.g., groundwater levels) are generated for each model, using the parameter values obtained from steps 2–6.

8. The model weights and variances of each ensemble member are calculated using the DREAM algorithm as explained in section 2.3.
9. The model weights are computed by summing the weights for all selected ensemble members of each conceptual model.
10. Finally, multi-model predictions are obtained by assessing predictive mean and variance using equations 7 and 8.

2.5 Alternative conceptual models

Hoge et al. (2019) concluded in their review paper that selection of alternative conceptual models is the most important aspect of Bayesian Model Averaging. Enemark et al. (2019) present a review of the conceptual hydrogeological model development. In our study, four alternative conceptual groundwater flow models have been selected from 15 possible alternative conceptual groundwater flow models. These initial 15 conceptual groundwater flow models were constructed using different geological interpretations and boundary conditions.

All alternative conceptual models were calibrated using observed groundwater level data for the same period. The performance of each model was evaluated based on different performance evaluation coefficients and information criterion statistics. Details about model development, calibration, evaluation and selection are provided in Mustafa et al. (2019). Obviously, the best option would be to use all 15 conceptual models. However, it would be computationally very expensive. Nevertheless, our main objective is not to predict the groundwater level of this study area. Rather our objectives are (i) to develop an integrated uncertainty quantification methodology that can quantify different sources of uncertainty of a groundwater flow model and thereby increase the reliability of the model prediction and (ii) the demonstration of the applicability of the proposed approach with real-world mode using simple personal computer. Therefore, the four best performing conceptual models where selected to reduce the computational effort in the Bayesian methodology. However, spatial heterogeneity of the aquifer properties is not considered as a part of conceptual model uncertainty. Peeters and Turnadge (2019) recommended based on their hypothetical setup that, for an aquifer with high recharge and high conductivity, spatial heterogeneity of the aquifer properties should be considered in developing a groundwater flow model. Hence, further studies could be conducted considering other alternative conceptualizations including spatial heterogeneity of the aquifer properties.

Later, the IBMUEF methodology has been implemented using the better performing four alternative conceptual models. The four selected alternative groundwater models are: (i) a one-layer model with boundary condition-5 (L1B5), (ii) a two-layer model with boundary condition-5 (L2B5), (iii) a two-layer models with boundary condition-4 (L2B4) and (iv) a three-layer models with boundary condition-5 (L3B5). Details about the selected conceptual models and model setup are explained in section 2.5.1 and 2.5.2.

2.5.1 Alternative conceptual models development

A cross sectional (A-A') view of the simplified hydrogeological models is shown in Figure 1d. First, three simplified alternative conceptual groundwater models were defined based on the geological stratification. The three models are a one-layered (L1), a two-layered (L2) and a three-layered (L3) model setup as shown in figure 1d. The bottom elevation of the aquifer in model was taken 50 m below sea level. In the one-layered model (L1), the whole model domain was considered as one hydro-stratigraphic unit and it was assumed that hydraulic properties are homogeneous and isotropic. The two-layered model (L2) consists of two layers where the average thickness of the top layer was 10 m (clay and loamy clay soil) and rest of the thickness was considered as the bottom layer. The model domain was divided into three different hydro-stratigraphic units to develop a three-layered model (L3). The top layer of the three-layered model was the same as for the two-layered model, but just below the top layer, a fine sand layer with an average thickness of 8 m was added in the three-layered model. The bottom layer of three-layered model consists of medium sand, coarse sand and coarse sand with gravel. Four or more layered models were not considered in this study because the information of the exact positions of the groundwater abstraction wells filter was unknown. Therefore, a further increase in layer numbers would increase the complexities of placing groundwater abstraction wells in the model domain.

One of the major factors that influences conceptual model uncertainty is related to the boundary conditions of the model (Wu & Zeng, 2013). Boundary conditions of groundwater models are often very uncertain, although the model results largely depend on these boundary conditions. A previous study in the Bengal basin observed that groundwater flows from north to south (Michael & Voss, 2009a, 2009b). On the other hand, there is a large wetland at the southeastern corner of the study area, as well as a large river (known as Ganges/Padma) within a few kilometers from the south boundary. Since exact boundary conditions were not known, five different potential sets of boundary conditions were conceptualized based on the above information. In this study, two sets of boundary conditions are used after an initial evaluation (Figure 2). Detailed description of the other boundary conditions and the evaluation procedure are explained in Mustafa et al. (2019). In boundary condition 4 (B4), a constant head boundary was considered on the north side of the model, where most of the river branches originate, assuming that groundwater flow direction is parallel to the river flow, and the southeastern part of the model, where a large wetland is located. At the south part of the model domain, a constant head is assigned because the great Ganges/Padma river is very near to the south boundary. In boundary condition 5 (B5), at the north and north-western boundary also at the south-eastern corner of the model domain, a constant head boundary was considered, based on the information that groundwater is flowing from north and northwestern to south in the study area (Michael & Voss, 2009a, 2009b). A constant head is assigned at the south-eastern corner of the model domain to represent the Chalan Beel wetland. The south and north-eastern boundaries are parallel to groundwater flow direction (Michael & Voss, 2009a, 2009b) hence no-flow boundaries are assigned at the south and north-eastern boundaries.

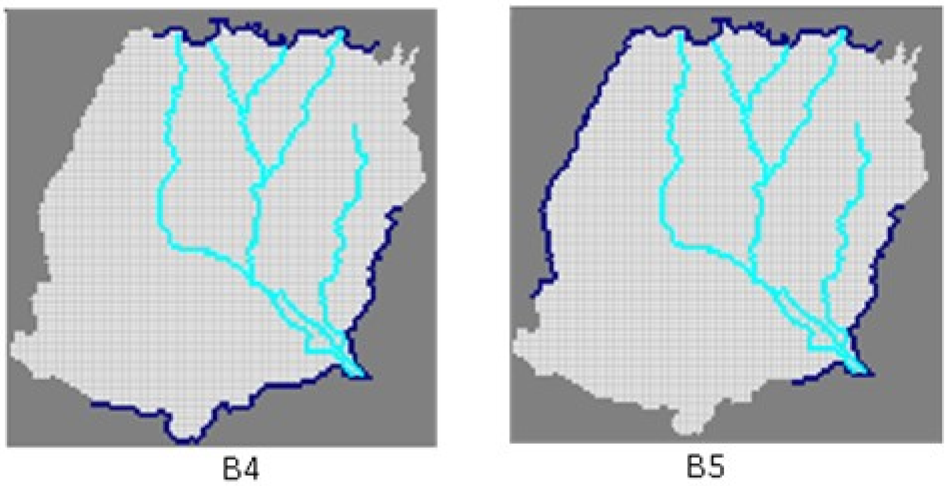


Figure 2: Alternative boundary conditions used to develop alternative conceptual model (blue line indicates constant head boundary): B4: constant head at north, south and southeast boundary; B5: constant head at north, northwestern and southeastern boundary.

2.5.2 Model setup and data

PMWIN: Processing MODFLOW (Chiang & Kinzelbach, 1998) is a grid based, fully-distributed, physically-based, integrated simulation system for modelling groundwater flow and solute transport processes and was used for groundwater flow simulations. The study area having an area of 7112 km^2 was discretized into smaller cells, resulting in 117 rows and 118 columns of grid cells, with a dimension of 900 m x 900 m. All the alterative conceptual models are transient with a monthly time step. A no-flow boundary is considered at the model domain bottom as vertical groundwater flow is restricted by the relatively impermeable hard rock below the aquifer in the study area. On the model top surface, a spatially distributed recharge boundary is considered. Spatially distributed monthly groundwater recharge was simulated using the WetSpass-M model with the same grid cell size as the MODFLOW model (Abdollahi et al., 2017; Batelaan & De Smedt, 2007). The study area was divided into 34 abstraction zone considering each upazila as one zone (upazila is the second lowest tier of regional administration in Bangladesh). Groundwater abstraction in each zone was calculated using an empirical relation based on the irrigated area and crop irrigation requirements. Details about the estimation of the groundwater abstraction and simulation of groundwater recharge can be found in Mustafa et al. (2017a).

The initial groundwater heads correspond to a long-term average groundwater table obtained by running the models in steady state conditions.

Weekly groundwater level and daily rainfall data were collected from the Water Resources Planning Organization (WARPO), Bangladesh. The groundwater level and rainfall were collected respectively for 140 and 30 sites. Available river discharge data of the BWDB for the existing small rivers within the study area were also collected from WARPO. Daily maximum and minimum temperature, wind speed and other climatic data were collected from the Bangladesh Meteorological Department (BMD). Reference evapotranspiration (ET_0) was calculated using the FAO Penman-Monteith equation (Allen et al., 1998; Mustafa et al., 2017a,b). In this study, reference evapotranspiration (ET_0) is also considered as potential evapotranspiration.

The monthly observed groundwater level data of 50 observation wells have been used for model calibration and validation (Figure 1b).

Topography and borehole data were collected from Bangladesh Multipurpose Development Authority (BMDA). The geological and lithological log data from twenty-three boreholes within the study area were collected from BMDA.

2.6 Parameterization

Groundwater recharge multipliers and groundwater abstraction multipliers have been introduced to quantify uncertainty of the estimated spatially distributed groundwater recharge and abstraction data. The input multipliers are considered as additional individual latent parameters during model calibration and uncertainty analysis and have been estimated along with model parameters. The hydrologically acceptable ranges of the multipliers have been defined based on the available knowledge of the possible level of bias in the initial estimation of groundwater recharge and abstraction (Table 1). In addition to the input multipliers, the following influential MODFLOW parameters have been considered: (i) Horizontal hydraulic conductivity, (ii) Specific yield, (iii) Hydraulic conductance of Riverbed and (iv) Specific storage. The first three MODFLOW parameters have been considered for the one-layered model. For the two- and three-layered models, specific storage has also been added. Considering specific parameters for each layer results in, respectively, seven and ten MODFLOW parameters to be considered for the two- and three-layered models (Table 1). The selected parameters and their prior uncertainty ranges are presented in Table 1.

A uniform prior probability distribution within the hydrologically acceptable ranges has been considered as a prior range for each parameter (Table 1) as we have no information about the distribution of the prior. Moreover, this is the most widely used prior in case of limited information availability about the distribution of the parameter value (Enemark et al. 2019). The range of hydrogeological parameter values was selected based on typical values for aquifer materials (Domenico & Mifflin, 1965; Domenico & Schwartz, 1998; Johnson, 1967) and previous research findings in the study area (Michael & Voss, 2009a, 2009b). Although the number of MODFLOW parameters is different for different conceptual model structures, the input multipliers and heteroscedastic error model parameters remain the same for all conceptual models (Table 1).

Table 1: Parameters of the alternative conceptual models, input multipliers and heteroscedastic error model parameters used in the uncertainty analysis using IBMUEF with their prior ranges

	Descriptions	**Unit**	**Ranges**
Input parameters for all models			
m_R	Groundwater recharge multipliers	-	0.010 – 10
m_A	Groundwater abstraction multipliers for temporal changes	-	0.010 – 10
The parameters of the heteroscedastic error model to consider heteroscedasticity of the groundwater level error			

A	Groundwater level uncertainty slope	-	0.010 – 1.0
B	Groundwater level uncertainty intercept	m	0.010 – 3.0
Model parameters of one-layer models (L1B5)			
HK	Horizontal hydraulic conductivity	m/s	0.0000001 – 0.0095
RIVC	Hydraulic conductance of Riverbed	m^2/s	0.001 – 1.6
SY	Specific yield	-	0.10 – 0.35
Model parameters of two-layer models (L2B5, L2B4)			
HK-1	Horizontal hydraulic conductivity of layer-1	m/s	0.0000001 – 0.0095
HK-2	Horizontal hydraulic conductivity of layer-2	m/s	0.0000001 – 0.0095
RIVC	Hydraulic conductance of Riverbed	m^2/s	0.001 – 1.6
SY-1	Specific yield of layer-1	-	0.10 – 0.35
SY-2	Specific yield of layer-2	-	0.10 – 0.35
SS-1	Specific storage multipliers of layer-1	-	0.015 – 15
SS-2	Specific storage multipliers of layer-2	-	0.015 – 15
Model parameters of three-layer models (L3B5)			
HK-1	Horizontal hydraulic conductivity of layer-1	m/s	0.0000001 – 0.0095
HK-2	Horizontal hydraulic conductivity of layer-2	m/s	0.0000001 – 0.0095
HK-3	Horizontal hydraulic conductivity of layer-3	m/s	0.0000001 – 0.0095
RIVC	Hydraulic conductance of Riverbed	m^2/s	0.001 – 1.6
SY-1	Specific yield of layer-1	-	0.10 – 0.35
SY-2	Specific yield of layer-2	-	0.10 – 0.35
SY-3	Specific yield of layer-3	-	0.10 – 0.35
SS-1	Specific storage multipliers of layer-1	-	0.015 – 15
SS-2	Specific storage multipliers of layer-2	-	0.015 – 15
SS-3	Specific storage multipliers of layer-3	-	0.015 – 15

2.7 Computational experiments

Three different scenarios have been used to perform uncertainty analysis along with model calibration. The model parameters and heteroscedasticity of groundwater level error have been considered in the first scenario. In this scenario, the input data are considered perfectly known and accurate. This scenario will serve as a benchmark. In the second scenario, model parameters, heteroscedasticity of the groundwater level error and temporal groundwater abstraction and recharge multipliers are considered. In this scenario, we introduced 12 groundwater recharge multipliers (m_R) to describe uncertainties in groundwater recharge, assigning a single multiplier corresponding to each time step which is one month in this study. Similarly, we introduced 6 groundwater abstraction multipliers (m_A) to describe uncertainties in groundwater abstraction, assigning a single multiplier corresponding to each time step. Abstraction multipliers have been considered only for the dry season (November to April), because this is the major abstraction period for irrigation in the study area. Details on estimation and uncertainty analysis of groundwater recharge and abstraction can be found in Mustafa et al. (2018).

Abstraction multipliers associated with the spatial estimation have been excluded in this study because of computational time although they might have considerable effect on the model prediction. In this study, four alternative conceptual groundwater models have been used with different levels of complexity. The computational time increases with increased complexity of the alternative conceptual groundwater models. For example, for the three-layer model with a total of 64 parameters (including both spatial and temporal abstraction multipliers), the algorithm has not reached convergence even after 200000 model evaluations. On a 2.70 GHz processor, 200000 model evaluations take around 21 days with an average of 9 seconds per simulation. Similarly, for the two-layered model with a total of 61 parameters (including both spatial and temporal abstraction multipliers), the algorithm has not been fully converged after 200000 model evaluations. This corresponds with around 19 days with an average of 8 seconds per simulation for the same processor. Of course, the evolution chain was converging towards the convergence both for the two and three-layered models. On the other hand, for the one-layered model with 57 parameters (including both spatial and temporal abstraction multipliers), the algorithm started to converge after 110000 model evaluations. Because of time limitations, abstraction multipliers associated with the spatial estimation have been excluded for all the alternative models in this study to have successful convergence results for all the models. However, we believe that this will not restrict the applicability of the approach because of the continuous advances in computational power.

Finally, in the third scenario, which we will refer to as IBMUEF in this study, conceptual model uncertainties are considered along with uncertainties from the model input, parameters and heteroscedasticity of groundwater level error. The IBMUEF framework is used to quantify all the mentioned sources of uncertainty in this scenario.

All the conceptual models have been calibrated and validated respectively for 1990 and 2000, for 12 monthly periods using 50 observation wells data for each period. It has been observed that models are able to accurately predict observation data which have not been used during the calibration. However, to ensure clear visualization, the results of 1990 are presented in the manuscript.

The d-factor, the % of observations within the 95 % confidence intervals (95% CI) and the Root Mean Square Error (RMSE) have been used to evaluate the model prediction uncertainty. The d-factor represents the average width of the 95% CI and is calculated as in (Yang et al., 2008):

$$\mathrm{d}-factor=\frac{\frac{1}{n}\sum_{t=1}^{n}\left(H_{t,u}-H_{t,l}\right)}{\sigma_0} \qquad (12)$$

Where $H_{t,u}$ and $H_{t,l}$ represent respectively, the upper and lower bounds of the 95% confidence intervals, n = the number of observations and σ_0 = the standard deviation of the observed groundwater level. d-factors closer to 1 indicate better model prediction (Yang et al., 2008). The higher observation coverage within the 95 % confidence intervals and decreasing d-factor value are indicating the improvement in model predictions and accuracy of the uncertainty bounds.

3. Results and discussion

In the results and discussion section, the results obtained from the three different scenarios as explained in the previous section (section 2.7) are presented, interpreted and discussed. Section 3.1 presents the parameter and prediction uncertainty of different conceptual models due to uncertainty of model parameters along with the heteroscedastic error model parameters. Section 3.2 elaborates on the parameter and prediction uncertainty of different conceptual models due to the uncertain input, model parameters along with the heteroscedastic error model parameters. Finally, section 3.3 presents the prediction uncertainty due to uncertainty of the conceptual model structure, input, model parameters and parameters of the heteroscedastic error model.

3.1 Parameter and prediction uncertainty of different conceptual models for scenario 1

Figure 3 shows the posterior probability distributions of the L1B5 model parameters for scenario 1. All parameters except riverbed hydraulic conductance (RIVC) of L1B5 model are well identified within their prior distribution. The posterior distribution of RIVC is still almost uniform while the posterior distribution of all other parameters is normally distributed, indicating that RIVC is a non-influential parameter. However, this could be improved in future studies by including more streamflow data during model calibration. We have also examined the correlation between model parameters and error model parameters. The results show a weak correlation among the MODFLOW parameters and between model parameters and error model parameters. The posterior distribution of SY is located at the lower boundaries of the prior range with a mean value of around 0.11. Alternatively, the posterior distribution of horizontal hydraulic conductivity (HK) is almost normally distributed with a high mean value of around 2.5 x 10^{-3} ms^{-1}. However, different conceptual models with different parameterization might draw different conclusions. Hence, consideration of conceptual model structural uncertainties may be important, but this is not considered in this scenario. Although the posterior probability distributions of the well identified parameters cover only a small range of their prior distributions, the parameter uncertainty band covers only 8.5% of the observations (Figure 5a). This can be argued as a problem of overconfidence in the estimation of the model parameters. Though the total uncertainty band covers almost all observations (94%), the width of the total uncertainty band is very wide compared to the width of the parameter uncertainty band. This is indicating that both the considered conceptual model structure and the input data used for this scenario contain a considerable amount of uncertainty.

Figure 4 shows the posterior pdfs of the L3B5 model parameters for scenario 1. As expected, the posterior parameter distributions of the L3B5 model are very different from the posterior parameter distributions of the L1B5 model. In this scenario, 12 parameters are considered, including two parameters of the heteroscedastic error model (A and B). Out of these 12 parameters, the posterior distributions of six parameters (HK-1, HK-2, HK-3, SY-1, a, and b) are approximately normally distributed. The posterior distribution of riverbed hydraulic conductance (RIVC) is still almost uniform like its prior distribution, again indicating that RIVC is a non-influential parameter. The posterior distributions of specific storage 1, 2 and 3 (SS-1, SS-2 and SS-3) are not included in the figure as the posterior distributions of those parameters are also still almost uniform as were their prior distributions. Similarly, the posterior distributions of specific storage for the two layered models also remain uniform, indicating that this is also a non-influential parameter (supplementary materials: Supplementary Figure 1). The posterior distributions of HK-1 and SY-2 are located respectively at the lower and upper boundaries of the prior range. Moreover, the posterior distribution of SY-3 is not well identified. This could be due to input uncertainties and/or conceptual model structural uncertainties which are not considered in this scenario. It also shows that the posterior probability distributions of the well identified parameters cover only a small range of their prior distributions except for HK-2. The parameter uncertainty band covers only 13 % of the observations (Figure 5d). Similar results are observed for the L2B4 and L2B5 models. For the L2B4 and L2B5 models, the parameter uncertainty band covers respectively 12 % and 13.8 % of the observations (Figure 5b, 5c and Supplementary Table 1). In general, the parameter uncertainty band is increasing with the level of complexity of the conceptual models. The observation coverage of the parameter uncertainty band for the different conceptual model structures is different. This suggests the importance of the use of multiple conceptual models for reliable prediction. Hoge et al. (2019) also suggested that consideration of uncertainty arising from conceptual physical interpretation is important during BMA implementation, if the objective of the study is to increase the reliability and accuracy of the model prediction.

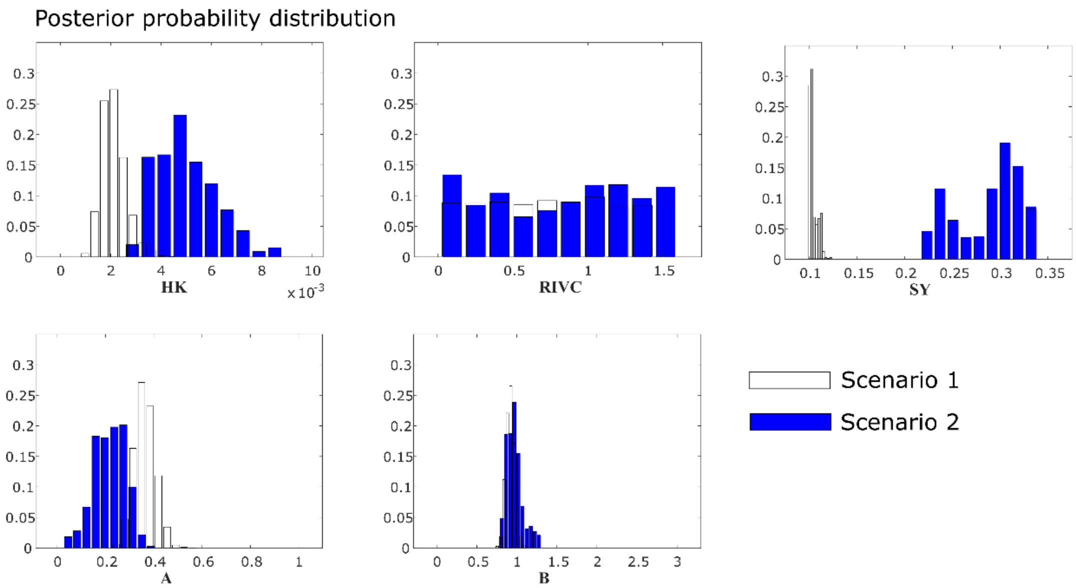
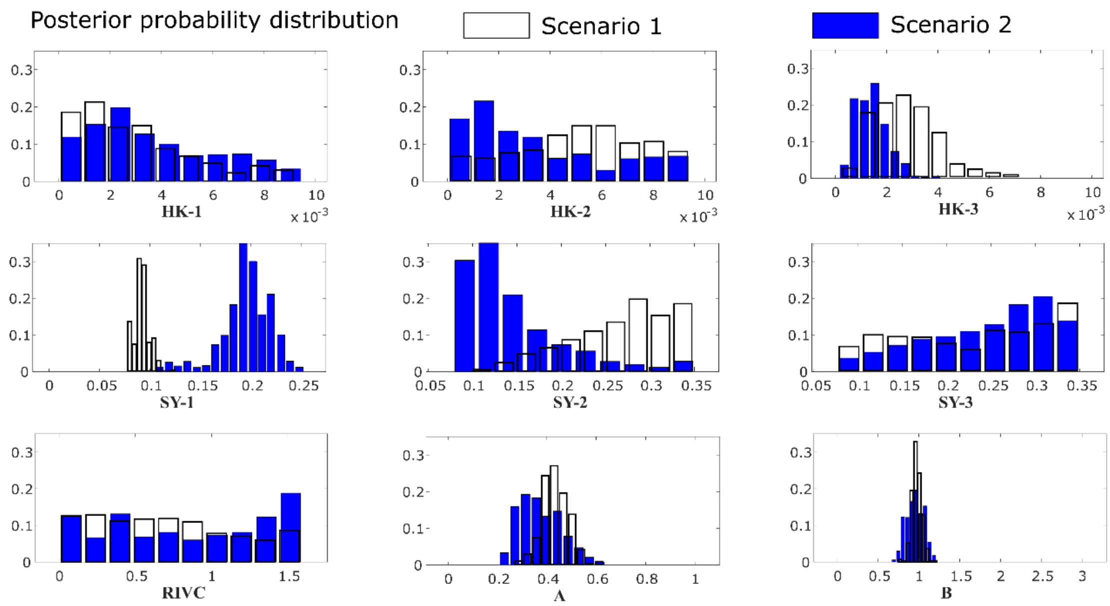


Figure 3: The posterior probability distribution of the L1B5 model parameters (top row) and the parameters of the heteroscedastic error-model (bottom row) both for scenario 1 and 2, using 2500 samples generated after convergence. HK: Horizontal hydraulic conductivity, RIVC: Hydraulic conductance of riverbed, SY: Specific yield, A and B: The parameters of the heteroscedastic error model.



662

663 Figure 4: The posterior probability distribution of the L3B5 model parameters and the
664 parameters of the heteroscedastic error-model (A and B) both for scenario 1 and 2, using
665 2500 samples generated after convergence.

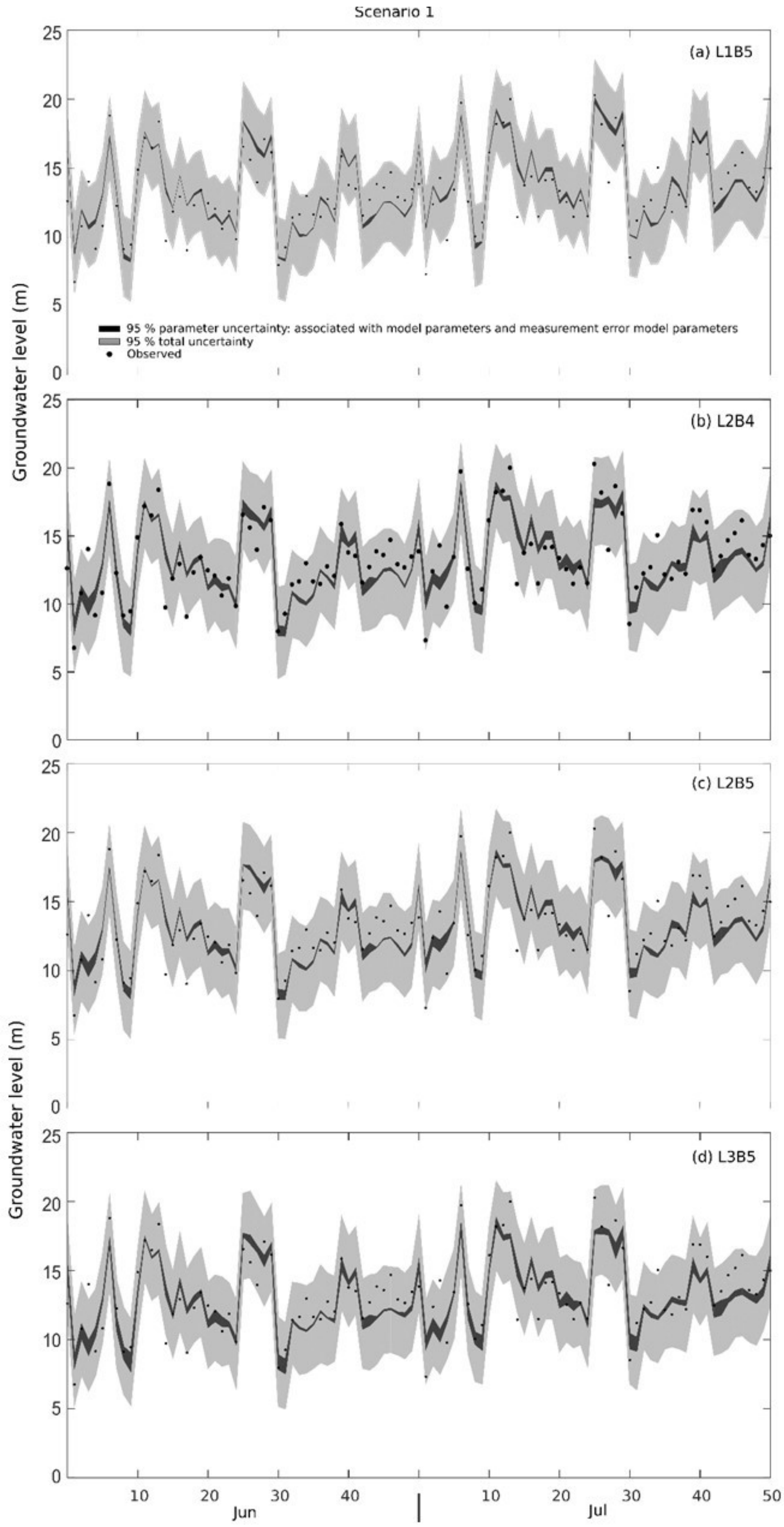


Figure 5: The prediction uncertainty of monthly groundwater level at each observation well with 95% parameter uncertainty considering error-model parameter along with model parameter (black interval), 95 % total uncertainty (dark gray) and observations (black dot) for (a) L1B5 model, (b) L2B4 model, (c) L2B5 model and (d) L3B5 model.

3.2 Parameter and prediction uncertainty of different conceptual models for scenario 2

In this scenario, uncertainty of the input data is quantified simultaneously along with model parameters and heteroscedastic error-model parameters.

Figure 3 shows the posterior pdfs of the L1B5 model parameters for scenario 2. As in scenario 1, all parameters are well identified within their prior ranges except RIVC. The posterior pdfs of the well identified parameters cover only a limited part of the prior range. The posterior distribution of the hydraulic conductance of riverbed (RIVC) is still almost uniform. Additionally, the posterior distribution of SY shows a slight multimodality. The correlation among model parameters and the correlation between model parameters, error model parameters and input multipliers have been examined. The results show a weak correlation among the MODFLOW parameters and between model parameters, error model parameters and input multipliers (recharge and abstraction multipliers).

Out of the 12 parameters for model L3B5, the posterior distributions of eight parameters (HK-1, HK-2, HK-3, SY-1, SY-2, SY-3, a, and b) are approximately normal while it was six for scenario 1 (Figure 4). The posterior distribution of RIVC, SS-1, SS-2, SS-3 are still almost uniform.

By comparing the posterior distributions between scenario 1 and 2 for different conceptual models (Figures 3 and 4), the following observations are made:

1. The posterior pdfs of some parameters are different in different conceptual models as well as in different scenarios. This is indicating that parameter values are overly adjusted to compensate for existing conceptual model structural deficiencies and input uncertainty when input and/or conceptual model uncertainties are not considered.
2. For model L3B5, the posterior pdfs of the parameters SY-2 and SY-3 are also identified within their prior ranges and their posterior distribution became approximately normal when we consider input uncertainty in addition to uncertainty arising from model parameters and heteroscedastic error model parameters. However, their posterior distributions are located at the boundaries of the prior range. This could be because of model structural uncertainties.
3. The heteroscedastic error model parameters (A and B) are well identified in both scenarios for all different conceptual models, but their values are different between

scenarios and between models. In general, the values of the error heteroscedasticity (A and B) parameters decrease when we consider input uncertainty in addition to uncertainty of model parameter and heteroscedastic error model parameters. Another important observation is that the value of the first error heteroscedasticity (A) parameter increases with the level of complexity of the conceptual models. This indicates that existing conceptual model structural deficiencies are somehow compensated by the value of the error heteroscedasticity (a) parameter.

We conclude that an explicit consideration of input uncertainty in addition to uncertainty of the model parameters and heteroscedastic error model parameters is very important to have unbiased and better defined parameter sets. Consideration of alternative conceptual models is also important for obtaining confident parameter sets. Schoniger et al. (2015) also reported that consideration of uncertainty arising from the model input is necessary to increase the robustness of Bayesian model averaging and ranking.

The posterior probability distributions of the recharge multipliers vary strongly between months, but are in general higher than one. The recharge multipliers are well identified during the rainy season (May to October), while these multipliers are not well identifiable during the dry season (November to April). The details of the recharge multipliers for a specific conceptual model are explained in Mustafa et al. (2018). The distributions of the well identified multipliers show different shapes for different conceptual models (Figure 6). However, the range of the multipliers and magnitude of their probability distributions are the same for different conceptual models (Figure 6). The groundwater abstraction multipliers are also well identified within their prior range and are higher than one in all months except for November and January for all four conceptual models. Again, the well identified multipliers show almost the same range of values for different conceptual models (Figure 7). This indicates that the input uncertainty multipliers are independent from model structural uncertainty and are not overly adjusted to compensate conceptual model structural deficiencies.

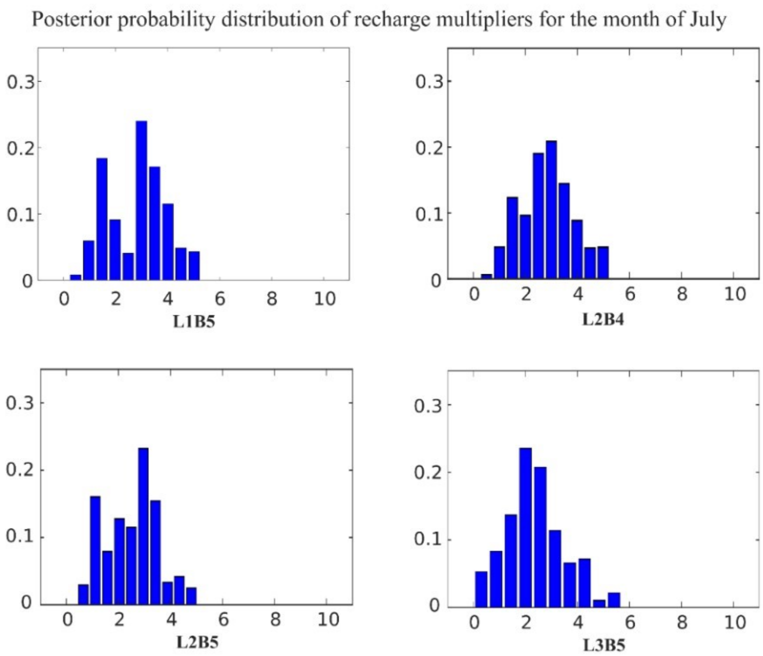


Figure 6: Posterior distribution of groundwater recharge multipliers of July for all conceptual models, using 2500 samples generated after convergence.

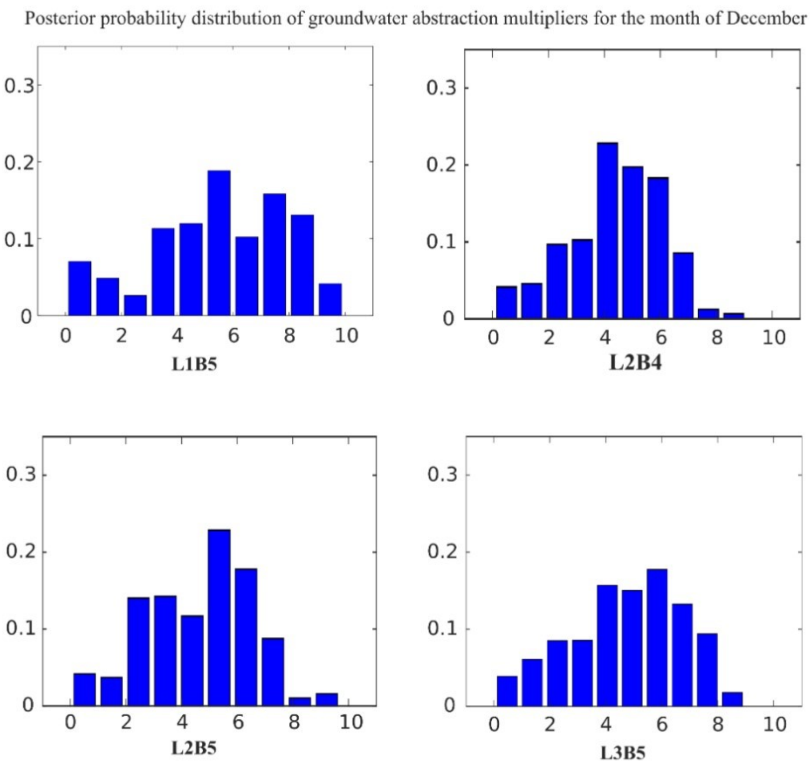


Figure 7: Posterior distribution of groundwater abstraction multipliers, using 2500 samples generated after convergence.

The prediction uncertainty of the monthly groundwater level associated with input uncertainty, model parameter uncertainty and uncertainty related to the heteroscedastic error model is presented in figure 8. The observation coverage of the parameter uncertainty band has increased by more than 100% for all models (Supplementary Table 1) when uncertainty arising from model input is incorporated along with uncertainty arising from model parameters and parameters of the heteroscedastic error model. The increase for the L1B5 model is even more than 200%. This result reveals that consideration of input uncertainty has significantly improved the confidence of model predictions and ignoring input uncertainty could lead to biased model simulations and incorrect uncertainty bands.. The parameter uncertainty band of L1B5 covers the highest number of observations when input uncertainty is included (Supplementary Table 1). When we explicitly consider input uncertainty, the width of the parameter uncertainty band has increased but the width of the total uncertainty has decreased (figure 5 and 8). This indicates that total uncertainty has decreased. This is confirmed by the reduction of the d-factor for all the models (Supplementary Table 1). This result reveals that uncertainty bounds of scenario 2 are more accurate compared to the CI of scenario 1, and the residual variance is smaller at each point. The Root Mean Square Error (RMSE) was also used to compare the results of scenario 1 and 2. It is observed that the values of the RMSE are decreasing when input uncertainty is included along with model parameter uncertainty and the parameters of the heteroscedastic error model (Figure 14). The decreasing magnitude of the RMSE value of L1B5 model is more significant than for any of the other models, indicating comparatively higher uncertainties in the L1B5 model. This is another indication that consideration of uncertainties through input multipliers is increasing the accuracy of the model prediction and decreasing the prediction uncertainty. Even after consideration of input uncertainties, the observation coverage of the parameter uncertainty band for the different conceptual model structures is different (Supplementary Table 1, Figure 8). Hence, consideration of conceptual model structural and input uncertainty is important to have more accurate model prediction and unbiased uncertainty bounds.

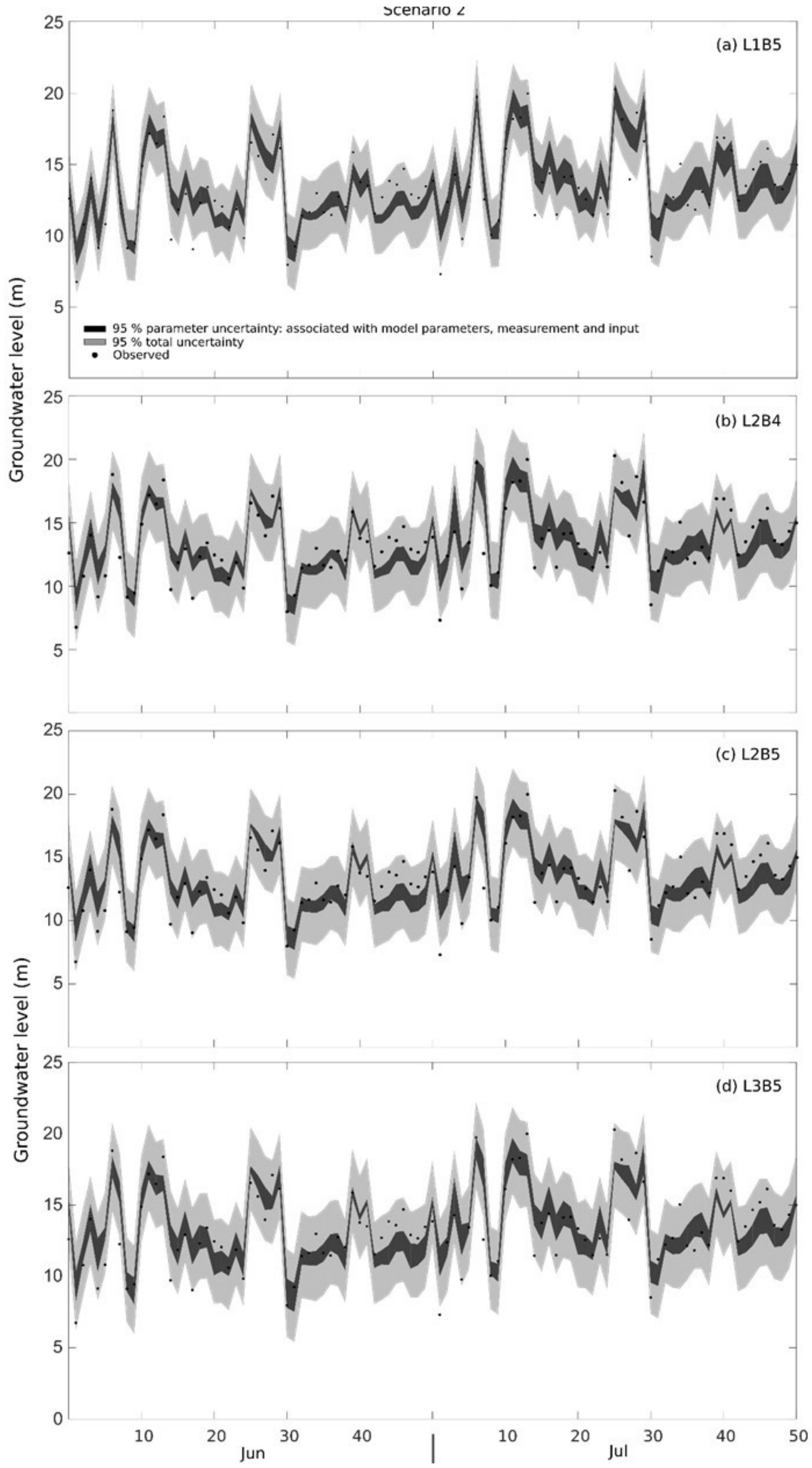


Figure 8: Prediction uncertainty of monthly groundwater level at each observation well with 95% parameter uncertainty considering input uncertainty along with model parameter uncertainty and error heteroscedasticity (black interval), 95 % total uncertainty (dark gray) and observation (black dot) for (a) L1B5 model, (b) L2B4 model, (c) L2B5 model and (d) L3B5 model.

3.3 Application of IBMUEF: assessment of the model uncertainty from input, model parameters, parameters of the heteroscedastic error-model and conceptual model structure

In the IBMUEF framework, uncertainties originating from the model input, the parameters, the parameters of the heteroscedastic error-model and the conceptual model structure can be taken into account. In this section, besides a presentation and discussion of the results of the IBMUEF application, these are also compared with the results of the previous scenarios.

Figure 9 shows parameter uncertainty bounds for all four alternatives conceptual models considering uncertainty arising from model input, parameter, and measurement heteroscedasticity. The different conceptual model structures cover different observations. This is indicating the skill of the models to capture different hydrogeological processes of the system.

The marginal densities of the estimated weights (following step 9 of section 2.4) for each model are shown in figure 10. The weight of the L1B5 model is well identified and has a normal distribution. Its likelihood value (weight) is very high compared to other models. The weight of all other models is very small. Nevertheless, their contribution is considered in the final results as they are representing different geological processes which are not considered in L1B5. The necessity of incorporating different models is also confirmed by the limited correlations between the groundwater level predictions using different conceptual models (Supplementary Table 2). For example, if a researcher/user knows the L1B5 model prediction, the L2B5 model adds more additional information to the final result compared to the L2B4 and L3B5 models as L2B5 is less correlated with L1B5. In general, correlations between the models are limited, indicating that different conceptual models are providing important information of the different hydrogeologic processes of the system. Hence, consideration of different conceptual models is needed to have a reliable model prediction.

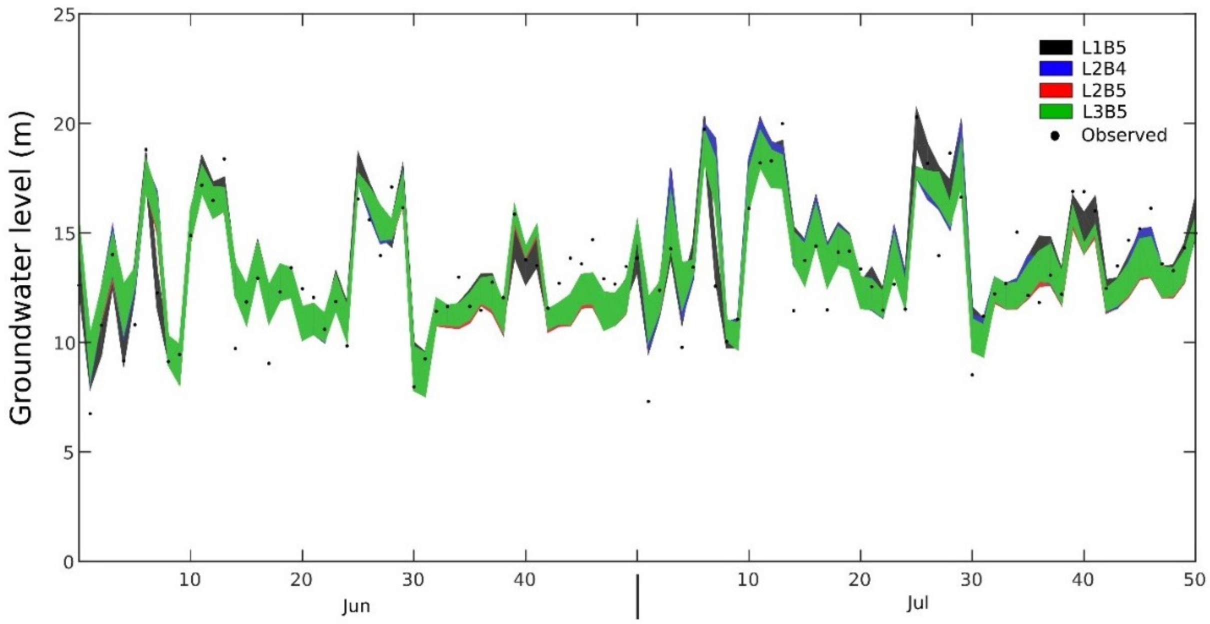


Figure 9: Prediction uncertainty of monthly groundwater level for all the conceptual models at each observation well with 95% parameter uncertainty considering input uncertainty along with model parameter uncertainty and error heteroscedasticity and observation (black dot).

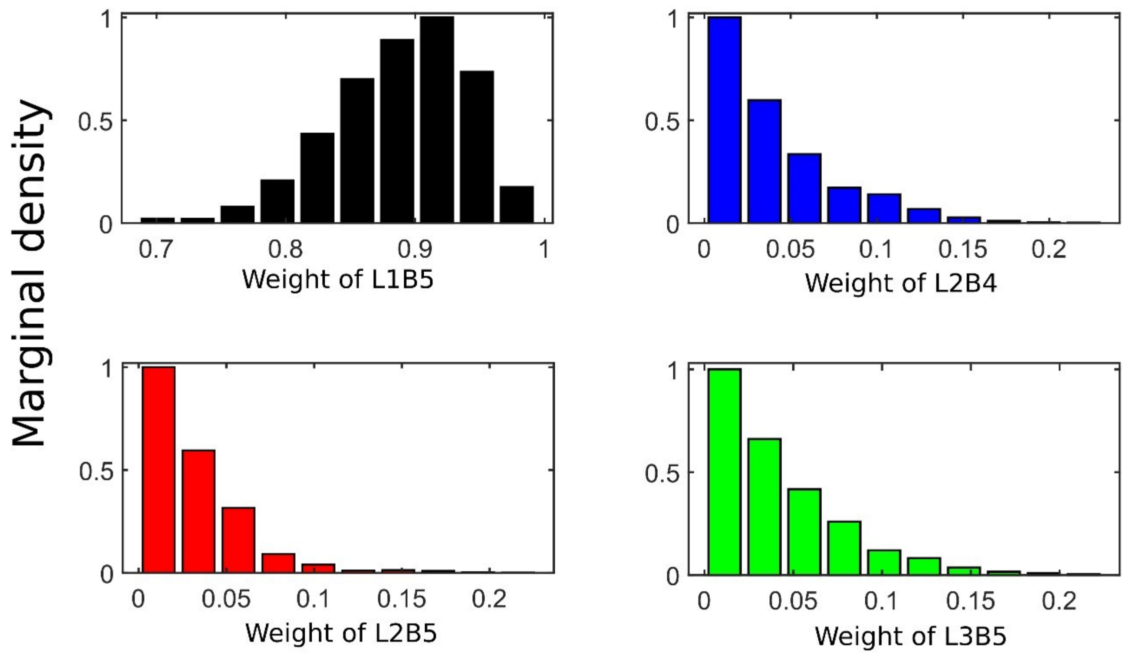


Figure 10: Marginal density of estimated weight for each model using integrated Bayesian multi-model uncertainty estimation framework (IBMUEF).

Figure 11 shows the final IBMUEF 95% prediction uncertainty of the monthly groundwater levels at each observation well considering model input, parameter, error heteroscedasticity model parameter, and conceptual model structural uncertainty. The final IBMUEF prediction was calculated using the prediction of the individual member models and their corresponding likelihood values (Figures 9 and 10) as explained in sections 2.3 and 2.4. The black line in figure 11 shows the mean prediction of the IBMUEF. The IBMUEF mean prediction and variance of figure 11 were calculated using equation 7 and 8, respectively. The distribution shape is determined by the weighted sum of the posterior distributions of each member model (Figure 12). It is observed that the posterior distribution of L1B5 model is capturing the reality more accurately compared to other models in the selected section of figure 11 (Figure 12). Hence, the distribution shape of the L1B5 model has a dominant role on the final IBMUEF prediction distribution shape of that section.

As expected, the 95 % CI of IBMUEF covers 95% of the observations which is very high compared to the individual models (Figure 11 and 13). Another interesting observation is that the d-factor value (1.42) has become smaller than the previous results. This is an indication of the improved model predictions and accuracy of the uncertainty bounds. The Root Mean Square Error (RMSE) was also used to evaluate the skill of the IBMUEF and to compare it with the individual model ensembles. The probability distributions of the RMSE-values for each of the models and IBMUEF are shown in figure 14. It is observed that the IBMUEF results in lower RMSE values compared to any individual model from the ensemble (Figure 14). This result reveals that the IBMUEF framework provides better model predictions. We conclude that an explicit consideration of conceptual model structural uncertainty is important for obtaining more accurate model predictions and unbiased uncertainty bounds. The results for this study are in line with results from similar approaches in surface water modeling (e.g., Ajami et al., 2007).

The IBMUEF framework is providing better and more reliable model predictions and more accurate uncertainty bounds, which is very important for decision support applications. However, as mentioned earlier, the implementation of the methodology is computationally expensive. The computational burden has also been identified as a main drawback for all other existing integrated uncertainty assessment approaches (Rojas et al., 2008; Ajami et al., 2007; Gelman et al., 2014). Based on their hypothetical setup, Xue & Zhang (2014) and

Hendricks Franssen et al. (2011) advocated that the EnKF is computationally more efficient compared to other existing Bayesian methods. However, comparison of the computational efficiency of the EnKF and other integrated Bayesian approaches with a real-world model remain unsolved. Another alternative could be Granger-Ramanathan averaging (GRA). GRA provides very similar performance as BMA, but is computationally less demanding (Diks and Vrugt, 2010). The information criterion (e. g.: AIC: Akaike information criterion) is another alternative to obtain a computationally less demanding approach (Hoge et al. 2019). However, model averaging based on AIC has been criticised by researchers as it is not based on a rigorous statistical basis and its results has no BMA interpretation (Wasserman, 2000; Tsai and Elshall, 2013). That's why it has been considered as a model selection technique instead of model averaging (Hoge et al. 2019). As a consequence, we have to choose between two different approaches: (i) computationally demanding but statistically robust, reliable and more accurate approaches or (ii) approaches without rigorous statistical foundation, which are computationally less demanding. Nonetheless, the statistically robust adaptive MCMC sampling of the DREAM algorithm is computationally more efficient for high-dimensional and multimodal application (Vrugt et al 2009a, 2016; Leta et al., 2015). Since the latter multi-chain MCMC based algorithm has been adopted in this study as a sampling approach, we believe that our approach is computationally more efficient compared to other existing integrated uncertainty assessment approaches. Moreover, the IBMUEF is a flexible framework as (i) there is no limitation for the number or complexity of alternative conceptual models and (ii) users can choose the number and dimensions (spatial and temporal) of input multipliers, based on the objectives of their modelling. It should be remembered that, the computational time increases with increases complexity of the alternative conceptual groundwater models and for a very complex model with more than 60 model parameters, the proposed approach became computationally very expensive. However, we believe that this will not restrict the applicability of the approach because of the continuous advances in computational power. Even though, effort should continue in the development of a more computationally efficient approach. We conclude that number or complexity of alternative conceptual models should be considered based on the modelling objectives during the implementation of a integrated uncertainty assessment approaches. Hoge et al. (2019) also concluded that the objective of the modelling should be the main driver in selecting model averaging approaches.

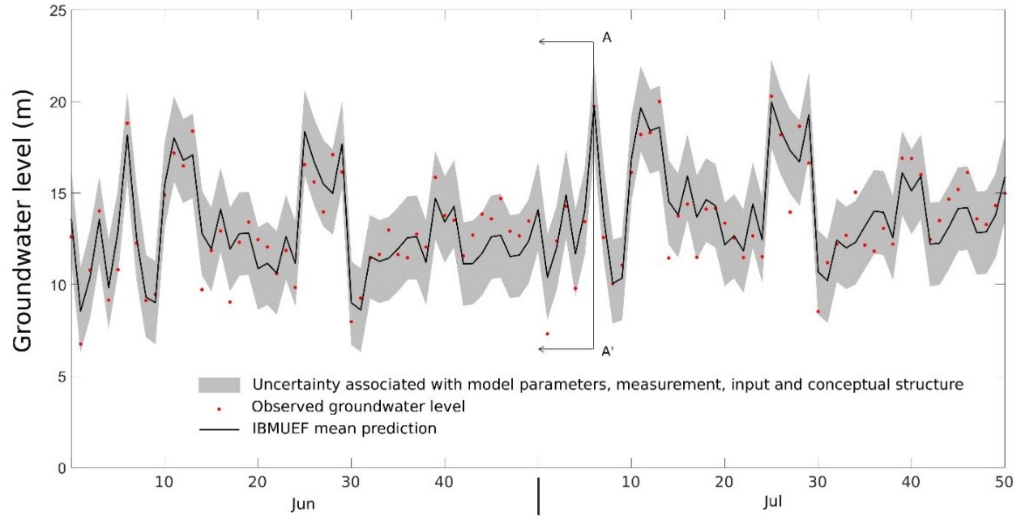


Figure 11: 95 % prediction uncertainty of monthly groundwater level at each observation well considering model input, parameter, error heteroscedasticity model parameter, and conceptual model structural uncertainty (gray shad), and IBMUEF predictive mean (black line), observation (red dot).

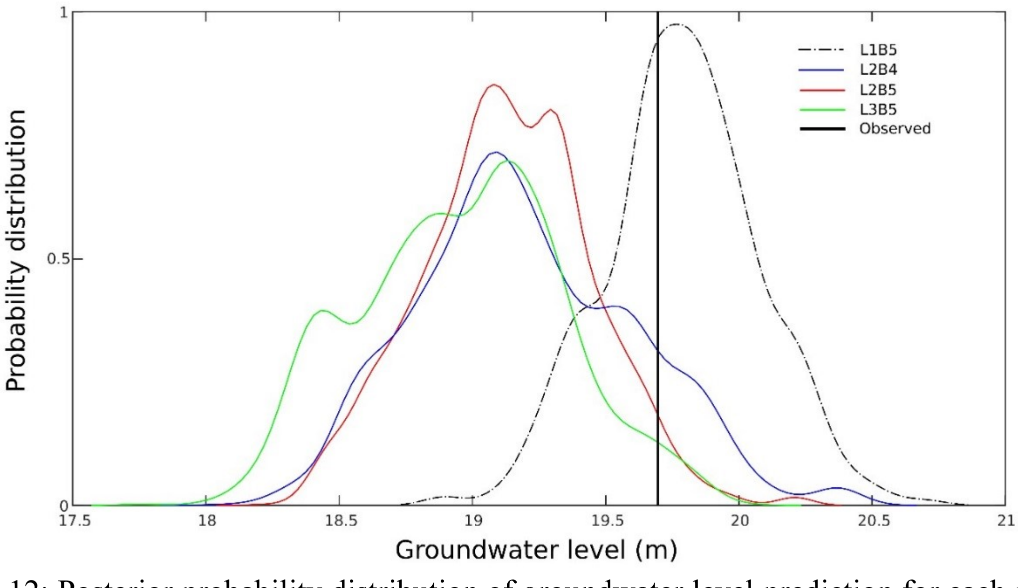


Figure 12: Posterior probability distribution of groundwater level prediction for each member model at the selected cross-section (A-A') of figure 11 and observed groundwater level (black line).

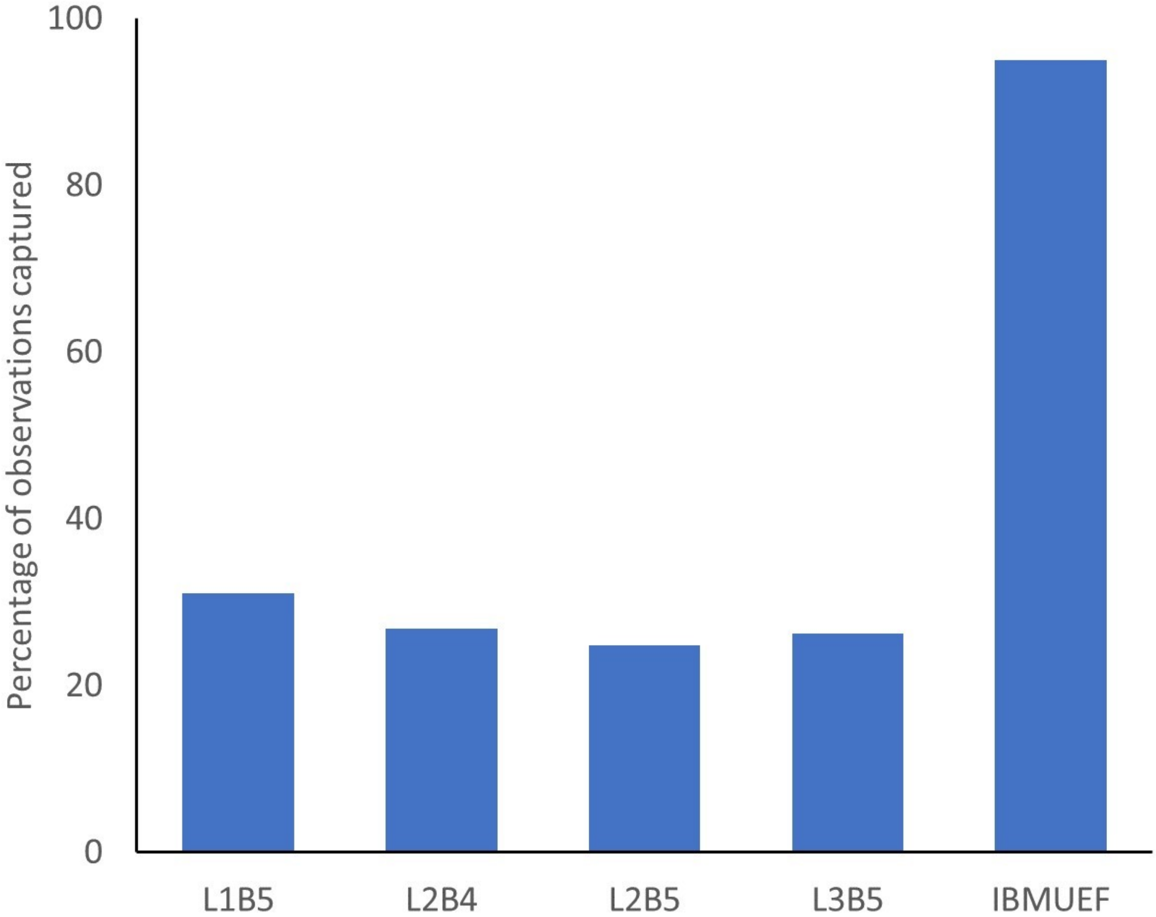


Figure 13: Percentage of observation captured by 95% parameter uncertainty bands of each conceptual model and IBMUEF.

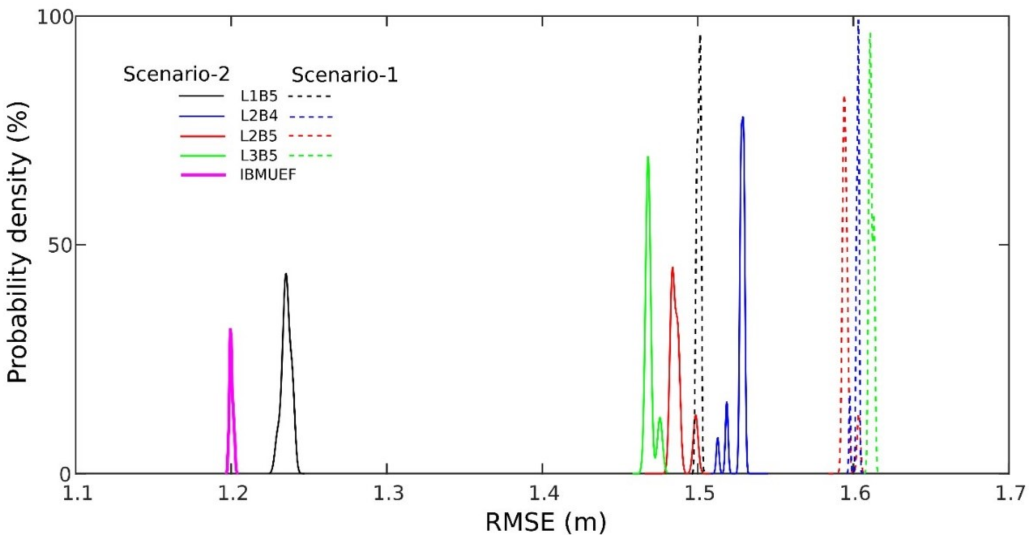


Figure 14: Probability distribution of RMSE for each model both for scenario 1 and 2 and IBMUEF.

4. Conclusions

We present an integrated Bayesian multi-model uncertainty estimation framework (IBMUEF) to explicitly quantify the uncertainty originating from errors in model conceptualization, input data, parameter values and measurement heteroscedasticity error of a fully distributed physically-based groundwater flow model. In the proposed integrated fully Bayesian multi-model framework, the DREAM algorithm with a specific likelihood function is combined with BMA. Groundwater recharge multipliers and groundwater abstraction multipliers are used in this framework to quantify uncertainty of spatially distributed input data of the groundwater model. The measurement heteroscedasticity is also considered in our integrated Bayesian framework by incorporating a novel heteroscedastic error model. To check the applicability of IBMUEF, four alternative conceptual models have been developed using a numerical groundwater flow model (MODFLOW) based on different interpretations of geological and hydrogeological information about the study area.

The results of this study confirm that conceptual model structure and uncertainty on the input data have a considerable effect on the model parameter distributions and model predictions. We demonstrated that parameter values are overly adjusted to compensate the existing conceptual model structural deficiencies and input uncertainty when they are not taken into account. Although consideration of input uncertainty results in better defined parameter distributions, consideration of alternatives conceptual models is also important to obtain confident parameter sets as the existing conceptual model structural deficiencies are somehow compensated by parameter uncertainties and the parameters of the heteroscedastic error model. On the other hand, input uncertainty multipliers appear to be independent from model structural uncertainty.

The total uncertainty of the system decreases but the observation coverage of the parameter uncertainty band increases by more than 100 % for all considered models when input uncertainty is included. Even when considering input uncertainty, the observation coverage of the parameter uncertainty band for the different conceptual model structures is different. This suggests the importance of the use of multiple conceptual models for reliable prediction. The parameter uncertainty band of L1B5 covers the highest number of observations when input uncertainty is included. This indicates that the L1B5 model is more capable of capturing the reality when input uncertainty is included. This is also confirmed by the highest likelihood (weight) value of the model. We demonstrate that consideration of input uncertainty along with model parameters uncertainty and measurement error generate more reliable model predictions. However, a very common limitation of these results is that the results are based on only a single conceptual model. Our results also confirm that even a very well calibrated conceptual model is unable to represent all the hydrogeologic processes of the system.

The IBMUEF prediction was calculated using the prediction of the individual member models and their corresponding likelihood values. The 95% prediction uncertainty band of IBMUEF covers 95 % of the observations which is significantly higher compared to any of the individual models. The IBMUEF framework has decreased the RMSE-value of the prediction and d-factor of the CI, and thereby increased the reliability of the prediction. The results of the study confirm that the IBMUEF framework is a useful tool to have better and more reliable model predictions and accurate uncertainty bounds. It is also shown that the IBMUEF is a useful and applicable framework to simultaneously quantify input, parameter, measurement and conceptual model uncertainty of a fully distributed physically-based groundwater flow model. We conclude that an explicit consideration of conceptual model structural uncertainty along with model input, parameter and measurement uncertainty using IBMUEF framework improves the accuracy and reliability of the model prediction and related uncertainty bounds.

Alternative conceptual models considered in this study have been developed using only MODFLOW. Future studies could be conducted considering different groundwater modelling algorithms to quantify the effect of numerical modelling errors. The modified log-likelihood function as explained in section 2.2, has been used in this study. However, future studies could be conducted considering different likelihood function to evaluate the effect of likelihood function.

In future studies, the framework can be implemented with more additional data sets to check the applicability with different prediction objectives e.g: baseflow. Moreover, application of the IBMUEF framework to quantify the groundwater level prediction uncertainties originating from the climate change and abstraction scenarios will increase the reliability of the model prediction and accuracy of the uncertainty bounds as its (IBMUEF) already consider all the other sources of uncertainties. However, number or complexity of alternative conceptual models or future scenarios should be considered based on the modelling objectives during implementation of this integrated uncertainty assessment approaches to avoid conceptual burden.

Acknowledgments

We thank Prof. Jasper Vrugt from University of California, Irvine, USA for his advice on the implementation of BMA. A draft version of a conference abstract appears online at AgEng2018.com but has not been published. The data used in this study are summarized and presented in the figures, tables, references and supporting information and will be available from the authors upon request (syed.mustafa@vub.be).

Author contributions

SM, JN and MH designed the study. SM and GG performed the analysis. SM and MH wrote the manuscript. All authors discussed the results and commented on the manuscript.

References

Abdollahi, K., Bashir, I., Verbeiren, B., Harouna, M. R., Van Griensven, A., Huysmans, M. & Batelaan, O. (2017). A distributed monthly water balance model: formulation and application on Black Volta Basin, *Environmental Earth Sciences*, 76(5), 198, https://doi.org/10.1007/s12665-017-6512-1.

Ajami, N. K., 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(1).

Akaike, H. (1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. *Annals of the Institute of Statistical Mathematics*, *26*(1), 363–387.

Batelaan, O., & De Smedt, F. (2007). GIS-based recharge estimation by coupling surface–subsurface water balances. Journal of Hydrology, 337(3-4), 337-355.

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

Beven, K., & Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298.

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

Chitsazan, N., & Tsai, F. T. C. (2015). A hierarchical Bayesian model averaging framework for groundwater prediction under uncertainty. *Groundwater, 53(2)*, 305-316.

Diks, C. G., & Vrugt, J. A. (2010). Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stochastic Environmental Research and Risk Assessment, 24(6)*, 809-820.

Doherty J. (2000), PEST - Model-independent parameter estimation. User's manual. Watermark Computing. Australia

Domenico, P. A., & Mifflin, M. D. (1965). Water from low-permeability sediments and land subsidence. *Water Resources Research*, *1*(4), 563–576.

Domenico, P. A., & Schwartz, F. W. (1998). *Physical and chemical hydrogeology* (Vol. 506). Wiley New York.

Draper, D. (1994). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, *56*.

Elshall, A. S., & Tsai, F. T. C. (2014). Constructive epistemic modeling of groundwater flow with geological structure and boundary condition uncertainty under the Bayesian paradigm. *Journal of Hydrology, 517*, 105-119.

Enemark, T., Peeters, L. J., Mallants, D., & Batelaan, O. (2019). Hydrogeological conceptual model building and testing: A review. *Journal of hydrology, 569*, 310-329.

Gaganis, P., & Smith, L. (2006). Evaluation of the uncertainty of groundwater model predictions associated with conceptual errors: A per-datum approach to model calibration. *Advances in Water Resources*, *29*(4), 503–514.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing, 24(6)*, 997-1016.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research, 48(8).*

Hendricks Franssen, H. J., Kaiser, H. P., Kuhlmann, U., Bauser, G., Stauffer, F., Müller, R., & Kinzelbach, W. (2011). Operational real-time modeling with ensemble Kalman filter of variably saturated subsurface flow including stream-aquifer interaction and parameter updating. *Water resources research, 47(2)*.

Hill, M. C., & Tiedeman, C. R. (2007). Effective groundwater model calibration: with analysis of data, sensitivities, predictions, and uncertainty. John Wiley & Sons.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 382–401.

Höge, M., Guthke, A., & Nowak, W. (2019). The Hydrologist's Guide to Bayesian Model Selection, Averaging and Combination. *Journal of Hydrology*, 572, 96-107.

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

Johnson, A. I. (1967). *Specific yield: compilation of specific yields for various materials*. US Government Printing Office.

Johnson, R. H. (2007). Ground water flow modeling with sensitivity analyses to guide field data collection in a mountain watershed. *Groundwater Monitoring & Remediation*, 27(1), 75-83.

Kavetski, D., Kuczera, G., & Franks, S. W. (2006a). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research, 42.

Kavetski, D., Kuczera, G., & Franks, S. W. (2006b). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. Water Resources Research, 42(3).

Kuczera, G., Kavetski, D., Franks, S., & 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), 161–177.

Laloy, E., Rogiers, B., Vrugt, J. A., Mallants, D., & Jacques, D. (2013). Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. *Water Resources Research*, 49(5), 2664-2682.

Leta, O. T., Nossent, J., Velez, C., Shrestha, N. K., van Griensven, A., & Bauwens, W. (2015). Assessment of the different sources of uncertainty in a SWAT model of the River Senne (Belgium). *Environmental Modelling & Software*, 68, 129-146.

Li, X., & Tsai, F. T.-C. (2009). Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod. *Water Resources Research*, *45*(9).

Madigan, D., Raftery, A. E., Volinsky, C., & Hoeting, J. (1996). Bayesianmodel averaging, in Proceedings of the AAAI Workshop on IntegratingMultiple Learned Models (pp. 77–83). AAAI Press, Portland, Oreg.

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

Michael, H. A., & Voss, C. I. (2009a). Controls on groundwater flow in the Bengal Basin of India and Bangladesh: regional modeling analysis. *Hydrogeology Journal*, *17*(7), 1561.

Michael, H. A., & Voss, C. I. (2009b). Estimation of regional-scale groundwater flow properties in the Bengal Basin of India and Bangladesh. *Hydrogeology Journal*, *17*(6), 1329–1346.

Minka, T. P. (2002). Bayesian model averaging is not model combination. Available electronically at http://www. stat. cmu. edu/minka/papers/bma. html, 1-2.

Montanari, A. (2005). Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, *41*(8).

Monteith, K., Carroll, J. L., Seppi, K., & Martinez, T. (2011, July). Turning Bayesian model averaging into Bayesian model combination. In *The 2011 International Joint Conference on Neural Networks (pp. 2657-2663)*. IEEE.

Mustafa, S. M. T., Nossent, J., Ghysels, G., & Huysmans, M. (2018). Estimation and Impact Assessment of Input and Parameter Uncertainty in Predicting Groundwater Flow with a Fully Distributed Model. *Water Resources Research*, 54(9), 6585-6608, doi: 10.1029/2017WR021857.

Mustafa, S. M. T., Abdollahi, K., Verbeiren, B., & Huysmans, M. (2017a). Identification of the influencing factors on groundwater drought and depletion in north-western Bangladesh. *Hydrogeology Journal*, 25(5), 1357–1375.

Mustafa, S. M. T., Vanuytrecht, E., & Huysmans, M. (2017b). Combined deficit irrigation and soil fertility management on different soil textures to improve wheat yield in drought-prone Bangladesh. *Agricultural Water Management*, 191, 124-137.

Mustafa, S. M. T., Hasan, M. M., Saha, A. K., Rannu, R. P., Van Uytven, E., Willems, P., & Huysmans, M. (2019). Multi-model approach to quantify groundwater level prediction uncertainty using an ensemble of global climate models and multiple abstraction scenarios. *Hydrology and Earth System Sciences,* 23(5), 2279-2303, https://doi.org/10.5194/hess-23-2279-2019.

Nettasana, T., Craig, J., & Tolson, B. (2012). Conceptual and numerical models for sustainable groundwater management in the Thaphra area, Chi River Basin, Thailand. *Hydrogeology Journal*, *20*(7), 1355–1374.

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

Peeters, L. J. M., & Turnadge, C. (2019). When to account for boundary conditions in estimating hydraulic properties from head observations?. *Groundwater, 57(3)*, 351-355.

Poeter, E., & Anderson, D. (2005). Multimodel ranking and inference in ground water modeling. *Groundwater*, *43*(4), 597–605.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174.

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

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process–a framework and guidance. *Environmental Modelling & Software*, *22*(11), 1543–1556.

Ridler, M. E., Zhang, D., Madsen, H., Kidmose, J., Refsgaard, J. C., & Jensen, K. H. (2018). Bias-aware data assimilation in integrated hydrological modelling. *Hydrology Research, 49(4),* 989-1004.

Rojas, R., Feyen, L., & Dassargues, A. (2008). Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research*, *44*(12).

Rojas, R., Kahunde, S., Peeters, L., Batelaan, O., Feyen, L., & Dassargues, A. (2010). Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling. *Journal of Hydrology*, *394*(3), 416–435.

Schöniger, A., Wöhling, T., Samaniego, L., & Nowak, W. (2014). Model selection on solid ground: Rigorous comparison of nine ways to evaluate B ayesian model evidence. *Water resources research*, 50(12), 9484-9513.

Schöniger, A., Wöhling, T., & Nowak, W. (2015). A statistical concept to assess the uncertainty in B ayesian model weights and its impact on model ranking. *Water Resources Research, 51(9),* 7524-7546.

Singh, A., Mishra, S., & Ruskauff, G. (2010). Model averaging techniques for quantifying conceptual model uncertainty. *Groundwater*, *48*(5), 701–715.

Stedinger, J. R., Vogel, R. M., Lee, S. U., & Batchelder, R. (2008). Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research*, *44*(12).

Troldborg, L., Refsgaard, J. C., Jensen, K. H., & Engesgaard, P. (2007). The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. *Hydrogeology Journal*, *15*(5), 843–860.

Troldborg, M., Nowak, W., Tuxen, N., Bjerg, P. L., Helmig, R., & Binning, P. J. (2010). Uncertainty evaluation of mass discharge estimates from a contaminated site using a fully Bayesian framework. *Water Resources Research*, *46*(12).

Tsai, F. T. C. (2010). Bayesian model averaging assessment on groundwater management under model structure uncertainty. *Stochastic Environmental Research and Risk Assessment, 24(6)*, 845-861.

Tsai, F. T. C., & Elshall, A. S. (2013). Hierarchical Bayesian model averaging for hydrostratigraphic modeling: Uncertainty segregation and comparative evaluation. *Water Resources Research, 49(9)*, 5520-5536.

Van Straten, G. T., & Keesman, K. J. (1991). Uncertainty propagation and speculation in projective forecasts of environmental change: A lake-eutrophication example. *Journal of Forecasting*, *10*(1-2), 163–190.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, 75, 273-316.

Vrugt, J. A. (2016a). *MODELAVG: A MATLAB Toolbox for Postprocessing of Model Ensembles* (Vol. Manual). Department of Civil and Environmental Engineering, University of California Irvine, 4130 Engineering Gateway, Irvine, CA.

Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, *43*(1).

Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, 44(12).

Vrugt, J. A., Ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., & Higdon, D. (2009a). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation, 10(3)*, 273-290.

Vrugt, J. A., Ter Braak, C. J., Gupta, H. V., & Robinson, B. A. (2009b). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?. *Stochastic environmental research and risk assessment, 23(7)*, 1011-1026.

Vrugt, J. A., ter Braak, C. J., Diks, C. G., & Schoups, G. (2013). Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. *Advances in Water Resources, 51*, 457-478.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of mathematical psychology, 44(1)*, 92-107.

Xue, L., & Zhang, D. (2014). A multimodel data assimilation framework via the ensemble Kalman filter. *Water Resources Research, 50(5)*, 4197-4219.

Yang, J., Reichert, P., Abbaspour, K. C., Xia, J., & Yang, H. (2008). Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. Journal of Hydrology, 358(1), 1–23.

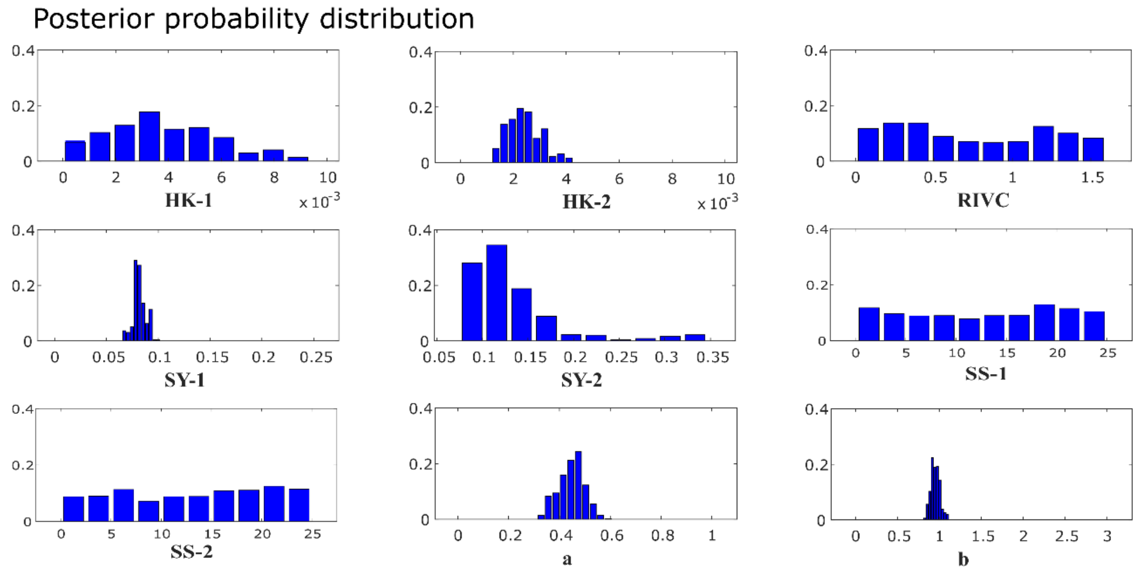
Ye, M., Neuman, S. P., & Meyer, P. D. (2004). Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research*, *40*(5).

Ye, M., Pohlmann, K. F., Chapman, J. B., Pohll, G. M., & Reeves, D. M. (2010). A model-averaging method for assessing groundwater conceptual model uncertainty. *Groundwater*, *48*(5), 716–728.

Yin, J., & Tsai, F. T. C. (2018). Saltwater scavenging optimization under surrogate uncertainty for a multi-aquifer system. *Journal of hydrology, 565*, 698-710.

Zhou, Y., & Herath, H. M. P. S. D. (2017). Evaluation of alternative conceptual models for groundwater modelling. *Geoscience Frontiers*, 8(3), 437-443.

Supplementary material



Supplementary Figure 2. The posterior probability distribution of the L2B4 model parameters and the parameters of the heteroscedastic error-model (A and B) for scenario 1, using 2500 samples generated after convergence.

Supplementary Table 3. Percentage observation coverage of the parameter uncertainty band and calculated d-factor based on the total uncertainty band for all the conceptual models.

	L1B5		L2B4		L2B5		L3B5	
	% cover	d-factor	% cover	d-factor	% cover	d-factor	% cover	d-factor
Scenario 1	8.5	1.88	12.0	2.03	13.8	2.01	13.0	2.04
Scenario 2	31.0	1.59	26.8	1.94	24.8	1.89	26.16	1.88

Supplementary Table 2. Correlations between the groundwater level predictions using different conceptual models.

	L1B5	L2B4	L2B5	L3B5
L1B5	1	-0.495	-0.367	-0.491
L2B4		1	-0.125	-0.119
L2B5			1	-0.072
L3B5				1