Bayesian Assessment of Conceptual Uncertainty in Hydrosystem Modeling

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Ing. Anneli Schöniger
aus Dresden

Tübingen
2016

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	03.02.2016
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr.-Ing. Olaf A. Cirpka
2. Berichterstatter:	Prof. Dr.-Ing. Wolfgang Nowak
3. Berichterstatter:	Prof. Dr. Walter A. Illman
4. Berichterstatter:	Prof. Dr. Frank Tsai

Abstract

This dissertation aims at improving uncertainty assessment for hydrosystem models subject to uncertainty in model structure, parameters, and forcing terms. In order to explicitly account for conceptual uncertainty (the uncertainty in model choice), Bayesian model averaging (BMA) is used as an integrated modeling framework. BMA is a formal statistical approach that rests on Bayesian probability theory. Weights are assigned to a set of alternative conceptual models based on their individual goodness-of-fit against observed data and the principle of parsimony. With these weights, model ranking, model selection or model averaging can be performed. The conceptual uncertainty within the set of considered models can be quantified as so-called between-model variance. A major obstacle to the wide-spread use of BMA lies in the computational challenge to evaluate BMA weights accurately and efficiently. The first part of this dissertation addresses this challenge by assessing and comparing different methods to evaluate the BMA equations, considering both mathematical approximations and numerical schemes (*Schöniger et al.*, 2014). Results of two synthetic test cases and of a hydrological case study show that the choice of evaluation method substantially influences the accuracy of the obtained weights and, consequently, the final model ranking and model-averaged results. If correctly evaluated, BMA weights point the modeler to an optimal trade-off between model performance and complexity. To determine which level of complexity can be justified by the available calibration data, the complexity component of the Bayesian trade-off is isolated from its performance counterpart in the second part of this dissertation. This model justifiability analysis (*Schöniger et al.*, 2015a) is demonstrated on model selection between groundwater models of vastly different complexity. The third part of this dissertation addresses the question of whether model weights are reliable under uncertain model input or calibration data. The proposed sensitivity analysis allows to assess the related confidence in model ranking (*Schöniger et al.*, 2015b). The impact of noisy calibration data on model ranking is investigated in an application to soil-plant model selection. Results show that model weights can be highly sensitive to the outcome of random measurement errors, which compromises the significance of model ranking. The findings from this dissertation also have important implications for the population and extension of the model set, for further model improvement, and for optimal design of experiments toward maximum confidence in model ranking. Overall, new statistical tools for model evaluation and uncertainty assessment are proposed, which are expected to be useful for a broad range of applications both in science and in practice.

Kurzfassung

Diese Dissertation hat zum Ziel, die Quantifizierung von Unsicherheiten bei der Modellierung von Hydrosystemen mit unsicherer Modellstruktur, unsicheren Parametern und unsicheren Eingangsdaten zu verbessern. Um explizit auch die Unsicherheit in der Modellwahl berücksichtigen zu können, wird Bayessche Modellmittelung (BMA) zur integralen Modellierung verwendet. BMA ist ein formaler statistischer Ansatz, der auf der Bayesschen Wahrscheinlichkeitstheorie beruht. Für ein Ensemble von alternativen Modellen werden Gewichte anhand der individuellen Kalibrierungsgüte und des Parsimonie-Prinzips bestimmt. Mit diesen Gewichten kann Modellranking, Modellwahl und Modellmittelung betrieben werden. Die konzeptionelle Unsicherheit innerhalb des Modellensembles kann als "zwischen-Modell-Varianz" quantifiziert werden. Ein großes Hindernis, das der weitverbreiteten Anwendung von BMA zur integrierten Modellierung und Unsicherheitsabschätzung im Wege steht, liegt in der technischen Herausforderung, BMA-Gewichte exakt und effizient zu bestimmen. Der erste Teil dieser Arbeit geht diese Herausforderung an mit einem Vergleich von verschiedenen Methoden zur Auswertung der BMA-Gleichungen unter Berücksichtigung sowohl mathematischer Annäherungen als auch numerischer Verfahren (*Schöniger et al.*, 2014). Die Ergebnisse zweier synthetischer Fallstudien und eines hydrologischen Anwendungsfalls zeigen, dass die Wahl des Auswerteverfahrens die Genauigkeit der ermittelten Gewichte wesentlich beeinflusst und damit auch das daraus folgende Modellranking und die modellgemittelten Ergebnisse. Sofern korrekt berechnet, zeigen die BMA-Gewichte einen optimalen Kompromiss zwischen Modellgüte und Komplexität auf. Um herauszufinden, welcher Komplexitätsgrad durch den vorhandenen Kalibrierungsdatensatz gerechtfertigt werden kann, wird im zweiten Teil der Arbeit die Komplexitätskomponente des Bayesschen Kompromisses von der Gütekomponente getrennt. Diese Modellrechtfertigungsanalyse (*Schöniger et al.*, 2015a) wird anhand der Modellwahl zwischen sehr unterschiedlich komplexen Grundwassermodellen demonstriert. Der dritte Teil der Arbeit befasst sich mit der Frage, ob die Modellgewichte unter unsicheren Modelleingangs- oder Kalibrierungsdaten zuverlässig sind. Die vorgeschlagene Sensitivitätsanalyse dient dazu, das zulässige Vertrauen in das resultierende Modellranking richtig einzuschätzen (*Schöniger et al.*, 2015b). Die Auswirkungen von verrauschten Kalibrierungsdaten auf das Modellranking werden anhand eines Fallbeispiels zur Boden-Pflanzen-Modellwahl untersucht. Die Ergebnisse zeigen, dass Modellgewichte sehr empfindlich auf den zufälligen Messfehler reagieren, was die Aussagekraft des Modellrankings beeinträchtigt. Die Erkenntnisse aus dieser Dissertation haben außerdem Bedeutung für die Auswahl und Erweiterung

des Modellensembles, für die Modellweiterentwicklung und für die optimale Datenerhebung im Sinne eines maximal zuverlässigen Modellrankings. Insgesamt werden neue statistische Instrumente zur Modellbewertung und Unsicherheitsanalyse vorgeschlagen, die für ein breites Anwendungsspektrum sowohl in der Wissenschaft als auch in der Praxis nützlich sein werden.

Acknowledgments

This dissertation would not have been the same, or even come into existence, without all the different kinds of encouragement along the way. My special thanks go out to...

- my supervisor Wolfgang Nowak for providing great support and inspiration, for offering guidance when necessary and freedom whenever possible, and for being a living proof that there does not have to be a tradeoff between speed and quality when giving feedback on manuscripts.
- my co-supervisor Thomas (Eddy) Wöhling for always taking the time for insightful discussions (both on- and off-topic) no matter how packed his schedule and for largely contributing to the success of this work with his expertise in modeling.
- my co-supervisor Walter A. Illman for his enthusiastic support of our collaborative case study and for hosting me at the University of Waterloo, Canada.
- the German Research Foundation (DFG) for funding within the international research training group "Integrated Hydrosystem Modelling".
- the head of the IRTG Olaf A. Cirpka for his unshakable faith in my research and time management, his valuable guidance and his inexhaustible motivation.
- all fellow doctoral students of the IRTG and my other housemates of Keplerstr. 17 as well as the hydrogeology working group and the administrative staff in Hölderlinstr. 12 for contributing to these memorable and pleasant last years.
- the first generation of the Wolf pack (Philipp Leube, Andreas Geiges, Felipe de Barros, Jonas Koch, Sergey Oladyshkin) for grand road trips, Caipirinha nights and Star Wars tales, and the second generation (Julian Mehne, Michael Sinsbeck, Felix Bode, Sebastian Most) for fabulous team building measures.
- Joanna McMillan, Jürnjakob Dugge, Maxi Herberich, Jonas Koch, Matthias Loschko and Michael Sinsbeck for additionally supporting me by reviewing parts of this dissertation.
- Uli Schollenberger, Stefan Spitzberg, Michael Boger and all my colleagues at BoSS Consult for having encouraged this endeavor from the first moment on.
- my friends for tolerating weird working hours and nerdy jokes, my parents for knowing best how to take the right decisions under uncertainty, and Philipp Guthke for his unconditional support no matter how creative the idea of my future career.

Herzlichen Dank!

Contents

List of Figures

List of Acronyms

AIC	Akaike information criterion
AICc	Bias-corrected AIC
BIC	Bayesian information criterion
BMA	Bayesian model averaging
BME	Bayesian model evidence
KIC	Kashyap information criterion
KIC@MAP	KIC evaluated at the maximum a posteriori parameter estimate
KIC@MLE	KIC evaluated at the maximum likelihood parameter estimate
MCMC	Markov chain Monte Carlo
TOM	Theoretically optimal model

List of Symbols

M	Model
$\mathcal{M}$	Model set
N_m	Number of alternative models
$\boldsymbol{\varphi}$	Model predictions
$\boldsymbol{u}$	Model parameters
$\boldsymbol{v}$	Model inputs
$\boldsymbol{\Theta}$	Uncertain model parameters and inputs
$\boldsymbol{e}$	Model structural errors
$\boldsymbol{\epsilon}$	Measurement errors
$\boldsymbol{y}_o$, $\boldsymbol{y}$	Observed and predicted data set, respectively
$p(\cdot)$	Prior probability density function
$p(\cdot\|\cdot)$	Posterior (conditional) probability density function
$E[\cdot]$	Expected value
$V[\cdot]$	Variance
$p(\boldsymbol{y}_o\|\boldsymbol{\Theta})$	Likelihood (as a function of $\boldsymbol{y}_o$)
$L(\boldsymbol{\Theta}\|\boldsymbol{y}_o)$	Likelihood (as a function of $\boldsymbol{\Theta}$)
$p(\boldsymbol{y}_o\|M_k)$	Bayesian model evidence
$P(M_k)$	Prior model weight
$P(M_k\|\boldsymbol{y}_o)$	Posterior model weight
$BF(M_k, M_l)$	Bayes factor
$\boldsymbol{\omega}$	Uncertain variable addressed by resampling

1 Introduction

Modeling and uncertainty assessment of hydrosystems Water resources are subject to diverse uses by humankind, such as drinking water supply, agricultural irrigation, energy production, as well as industrial and household purposes. These uses are threatened by water scarcity and by water pollution. Therefore, the use and protection of water resources needs to be efficiently managed. Further, specific risks emerging from hydrosystems (e.g., flooding of inhabited areas) need to be anticipated and mitigated.

Numerical hydrosystem models help to guide such management decisions. Models of surface and subsurface water flow and solute transport through hydrosystems enhance our understanding of the natural system. Relevant threats to, or emerging from, the hydrosystem can be more reliably identified. Further, simulation models can be used to predict the hydrosystem's response to future stresses or potential outcomes of management actions. This allows for a systematic risk assessment as a basis for rational decision making (*Goodarzi et al.*, 2013).

Process-based models approximate the natural system with simplified conceptualizations of governing processes and with effective laws and parameters. Model parameters typically cannot be exactly measured for several reasons: first, measurement data are subject to measurement errors. Second, especially in subsurface hydrology, measurement data are sparse. And third, effective parameters might not be observable at the scale of interest. The *conceptual uncertainty* of how to adequately represent the relevant physical processes, the *parameter uncertainty*, and the *measurement uncertainty* introduce uncertainty into the derived model predictions. Further, model *input uncertainty* can arise due to noise in the observations, due to uncertainty in upscaling or downscaling of input data, or when the model is used for forecasting under future conditions.

Assessing and dissecting these four main sources of uncertainty (*Renard et al.*, 2010) for hydrological model predictions is of crucial importance for drawing the right conclusions and for making informed decisions. Besides the "scientific desire to accompany predictions with uncertainty estimates" (*Liu and Gupta*, 2007), there is a growing interest in confidence intervals of predictions in practice. This is due to the fact that, depending on the risk aversion of a decision maker, the preference ranking of alternative management options might change if these options differ in their robustness against uncertainty (*Hall and Solomatine*, 2008).

Much effort has been made in the last decades to quantify parameter uncertainty in hydro(geo)logical models (e.g., *Kitanidis*, 1986; *Beven and Binley*, 1992; *Hill and Tiedeman*, 2007; *Gallagher and Doherty*, 2007; *Nowak*, 2010; *Liu et al.*, 2010). The treatment of input uncertainty is especially relevant in the field of surface hydrology due to the large variability of precipitation in time and space (e.g., *Kavetski et al.*, 2006a,b; *Vrugt et al.*, 2008).

Also the uncertainty in model choice itself, i.e., the conceptual uncertainty, has been recognized as an "integral part of inference" (*Buckland et al.*, 1997) by researchers of various disciplines (e.g., *Burnham and Anderson*, 2003; *Murphy et al.*, 2004; *Refsgaard et al.*, 2006; *Rojas et al.*, 2008; *Clark et al.*, 2011). Especially in coupled hydrosystem modeling, considering multiple model hypotheses can provide a much more realistic estimate of the overall predictive uncertainty. Not only input and parameter uncertainty of individual models can be assessed, but also the conceptual uncertainty of how to choose the most adequate representation of the (sub)system can be approximated. The latter would be neglected in single-model approaches, leading to a potentially severe underestimation of total predictive uncertainty. Figure 1.1 gives an overview of the different sources of uncertainty typically considered in hydrosystem modeling.

Multi-model approaches to account for conceptual uncertainty The proposition to retain a set of plausible theories (or alternative models) goes back to the foundations of the philosophy of science and is known as *Epicurus' principle of multiple explanations* (see e.g., *Hutter*, 2006). Following this philosophy, *Chamberlin* (1890) argued that keeping multiple hypotheses helps to avoid "the dangers of parental affection for a favorite theory". Reporting the uncertainty of model choice between several alternatives allows for a more robust decision making.

In such multi-model frameworks, the individual models are evaluated and ranked against each other. The ranking in form of model weights is based on predictive skill and, often, on some penalty for complexity such that more robust models are favored. A model-averaged estimate can be obtained from weighting the individual model predictions or statistics thereof. The conceptual uncertainty in the current set of models can be quantified as between-model variance.

The Bayesian multi-model approach The formal statistical approach of Bayesian model averaging (BMA) (*Draper*, 1995; *Hoeting et al.*, 1999) is a multi-model framework that rests on Bayesian probability theory. It combines *Epicurus' principle of multiple*

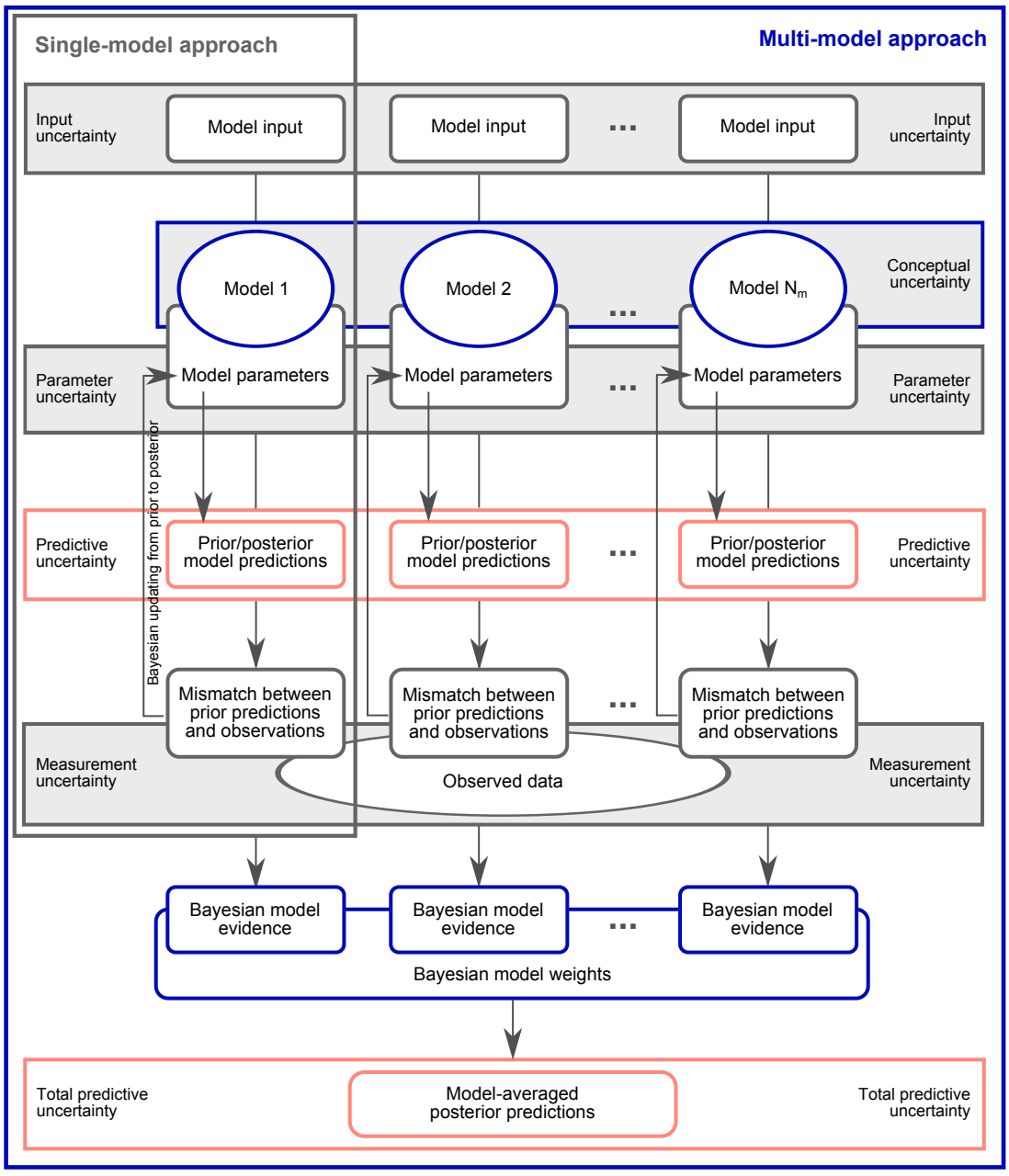


Figure 1.1: Illustration of the Bayesian approach to integrated modeling and uncertainty assessment of hydrosystem models, with the single-model approach framed in gray and Bayesian model averaging framed in blue. Sources of uncertainty are shaded in gray.

explanations (*Asmis*, 1984) with *Occam's razor* or *the principle of parsimony* (*Jefferys and Berger*, 1992): if competing models produce equally likely predictions, a higher model weight is assigned to the simpler explanation. I have chosen this framework to perform uncertainty analysis and to quantitatively assess the plausibility of hydrosystem models for its various advantages over other multi-model frameworks. The motivation for the use of BMA in this thesis is explained in Chapter 2.

BMA has been applied in various fields of research as a framework for model averaging (e.g., *Ajami and Gu*, 2010; *Najafi et al.*, 2011; *Seifert et al.*, 2012), model selection (e.g., *Raftery*, 1995; *Huelsenbeck et al.*, 2004), quantification of conceptual uncertainty (e.g., *Rojas et al.*, 2008; *Singh et al.*, 2010; *Troldborg et al.*, 2010; *Ye et al.*, 2010), data worth analysis (e.g., *Rojas et al.*, 2010; *Neuman et al.*, 2012; *Xue et al.*, 2014; *Wöhling et al.*, 2015), and model component dissection (*Tsai and Elshall*, 2013; *Elshall and Tsai*, 2014). The fields of application covered by the cited studies include sociology, ecology, and hydrology. In the course of this PhD project, I have further transferred the BMA methodology to the field of thermodynamic modeling (*Lötgering-Lin et al.*, 2015, in preparation).

Integrated approaches to modeling and uncertainty assessment Finally, there is a need to combine the efforts of assessing the individual sources of uncertainty for hydrological model predictions into an integrated modeling framework. Such a unified framework facilitates the development, the evaluation and the improvement of models. Scientists and practitioners benefit from an integrated approach by gaining more conceptual insight into the hydrosystem under study, by learning about model structural deficits, and by obtaining more reliable prognoses.

In the field of hydrology, two distinct approaches have been proposed to date: the Bayesian total error analysis (BATEA) (*Kavetski et al.*, 2006a; *Kuczera et al.*, 2006) and the integrated Bayesian uncertainty estimator (IBUNE) (*Ajami et al.*, 2007). BATEA quantifies the confidence in model predictions by considering all four mentioned sources of uncertainty: parameter uncertainty, measurement uncertainty, input uncertainty, and conceptual uncertainty. However, it does not pursue the idea of model ranking or multi-model combination. The modeler thus misses the opportunity to learn from a diagnostic model comparison as possible in a multi-model framework. The IBUNE approach allows to combine the predictions of multiple models, but uses a non-Bayesian optimization algorithm to obtain model weights. In this case, the modeler misses the advantage of a rigorous Bayesian derivation of posterior model probabilities.

In the field of hydrogeology, *Troldborg et al.* (2010) have demonstrated the usefulness of a fully Bayesian framework to account for uncertainty in geological model structure, in model parameters, and in measurements. While suggesting that BMA is a suitable framework to host the various components of uncertainty assessment, *Troldborg et al.* (2010) have pointed towards challenges with regard to the numerical implementation. From the previous studies cited here, it can be concluded that BMA is a promising approach for integrated modeling and uncertainty assessment of hydrosystems; however, there is a need for a systematic investigation of different aspects of BMA before a modeler can truly apply this framework with confidence.

Objectives and structure of the thesis This thesis aims at resolving selected issues that might hinder an efficient and meaningful use of BMA. The state of the art in Bayesian multi-modeling is presented in Chapter 2. Chapter 3 states the research questions addressed by this thesis. In Chapter 4, the developed concepts are presented and results from applying these concepts to different case studies of hydrosystem modeling are discussed. The corresponding publications can be found in Appendix A. In Chapter 5, I draw conclusions from this work and give an outlook toward potential future research.

2 State of the Art

The statistical framework of BMA is presented in Section 2.1. In Section 2.2, the ubiquitous problem of how to define the model set is touched upon. Further, the definition of the likelihood function required to perform Bayesian updating is discussed in Section 2.3. Section 2.4 focuses on the mechanism of the Bayesian tradeoff between performance and parsimony, which distinguishes BMA from other multi-model approaches. Challenges in the technical implementation of BMA are discussed in Section 2.5. In Section 2.6, the question of robustness of model weights against different sources of uncertainty is raised.

2.1 Statistical Framework of Bayesian Model Averaging

Consider a model M which yields model predictions $\boldsymbol{\varphi}$ as a function of $\boldsymbol{\Theta}$ and $\boldsymbol{c}$:

$$\boldsymbol{\varphi} = M(\boldsymbol{\Theta}) = f(\boldsymbol{\Theta}, \boldsymbol{c}), \tag{2.1}$$

with $\boldsymbol{\Theta}$ consisting of uncertain model parameters $\boldsymbol{u}$, and potentially uncertain model input $\boldsymbol{v}$, stochastic noise $\boldsymbol{w}$ (aleatory uncertainty), and model structural errors $\boldsymbol{e}$ (epistemic uncertainty), according to the modeler's conceptualization of the system under study. Prior knowledge about these variables can be formulated as probability density functions $p(\boldsymbol{u})$, $p(\boldsymbol{v})$, $p(\boldsymbol{w})$, and $p(\boldsymbol{e})$. Model input can refer to time-variant or constant forcings or boundary conditions. Model predictions $\boldsymbol{\varphi}$ might further depend on fixed input values or non-adjustable parameters, represented by $\boldsymbol{c}$. As soon as non-deterministic components $\boldsymbol{\Theta}$ are considered, a predictive distribution $p(\boldsymbol{\varphi})$ is obtained instead of deterministic predictions $\boldsymbol{\varphi}$.

Bayesian updating The prior probability density function of $\boldsymbol{\Theta}$ is updated to the posterior $p(\boldsymbol{\Theta}|\boldsymbol{y}_o)$ in light of the evidence in the observed data set $\boldsymbol{y}_o$ via Bayes' theorem:

$$p(\boldsymbol{\Theta}|\boldsymbol{y}_o) = \frac{p(\boldsymbol{y}_o|\boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{p(\boldsymbol{y}_o)} \propto p(\boldsymbol{y}_o|\boldsymbol{\Theta})\, p(\boldsymbol{\Theta}), \tag{2.2}$$

with $p(\cdot|\cdot)$ representing a conditional probability density function. $p(\boldsymbol{y}_o|\boldsymbol{\Theta})$ is the likelihood of a random realization from $p(\boldsymbol{\Theta})$ to have generated the observed data set.

The posterior predictive distribution $p(\boldsymbol{\varphi}|\boldsymbol{y}_o)$ is obtained from model runs based on $p(\boldsymbol{\Theta}|\boldsymbol{y}_o)$. The expected value of the prior or posterior predictive distribution of model M is denoted as $E[\boldsymbol{\varphi}]$ or $E[\boldsymbol{\varphi}|\boldsymbol{y}_o]$, respectively, and the variances are denoted as $V[\boldsymbol{\varphi}]$ and $V[\boldsymbol{\varphi}|\boldsymbol{y}_o]$, respectively.

This Bayesian updating or conditioning step corresponds to the calibration procedure in deterministic modeling applications. While Bayesian approaches to model calibration (or model comparison) are sometimes criticized for the need to specify prior beliefs, this is at the same time the beauty of Bayesian statistics: it forces modelers to make their assumptions transparent. Note that the specification of prior beliefs and of the likelihood function $p(\boldsymbol{y}_o|\boldsymbol{\Theta})$ are actually the only two definitions required to perform Bayesian updating. There are no limiting assumptions necessary on the structure of the models (e.g., linearity) or the shape of the involved prior parameter distributions (e.g., Gaussianity).

Bayesian model averaging Now consider a set $\mathcal{M}$ of N_m competing conceptual models M_k, with $k = 1 \dots N_m$. Applying BMA, the model-averaged posterior predictive distribution is determined as a linear mixture of the individual distributions:

$$p(\boldsymbol{\varphi}|\boldsymbol{y}_o) = \sum_{k=1}^{N_m} p(\boldsymbol{\varphi}|\boldsymbol{y}_o, M_k)\, P(M_k|\boldsymbol{y}_o)\,, \tag{2.3}$$

with model weights $P(M_k|\boldsymbol{y}_o)$. Formulating the weighted averaging of the individual models' statistics as the expected value over the model set

$$E_{\mathcal{M}|\boldsymbol{y}_o}[\cdot] = \sum_{k=1}^{N_m} [\cdot]\, P(M_k|\boldsymbol{y}_o), \tag{2.4}$$

yields an alternative expression for Equation 2.3:

$$p(\boldsymbol{\varphi}|\boldsymbol{y}_o) = E_{\mathcal{M}|\boldsymbol{y}_o}[p(\boldsymbol{\varphi})]\,. \tag{2.5}$$

The expected value of the posterior predictive distribution is determined as

$$\begin{aligned} E_{\boldsymbol{\varphi}|\boldsymbol{y}_o}[\boldsymbol{\varphi}] = & \sum_{k=1}^{N_m} E[\boldsymbol{\varphi}|\boldsymbol{y}_o, M_k]\, P(M_k|\boldsymbol{y}_o) \\ = & E_{\mathcal{M}|\boldsymbol{y}_o}\{E_{\boldsymbol{\varphi}|\boldsymbol{y}_o,\mathcal{M}}[\boldsymbol{\varphi}]\}, \end{aligned} \tag{2.6}$$

and the posterior variance as

$$\begin{aligned} V_{\boldsymbol{\varphi}|\boldsymbol{y}_o}\left[\boldsymbol{\varphi}\right] = & \sum_{k=1}^{N_m} V\left[\boldsymbol{\varphi}|\boldsymbol{y}_o, M_k\right] P\left(M_k|\boldsymbol{y}_o\right) \\ & + \sum_{k=1}^{N_m} \left(E\left[\boldsymbol{\varphi}|\boldsymbol{y}_o, M_k\right] - E\left[\boldsymbol{\varphi}|\boldsymbol{y}_o\right]\right)^2 P\left(M_k|\boldsymbol{y}_o\right) \\ = & E_{\mathcal{M}|\boldsymbol{y}_o}\{V_{\boldsymbol{\varphi}|\boldsymbol{y}_o,\mathcal{M}}\left[\boldsymbol{\varphi}\right]\} + V_{\mathcal{M}|\boldsymbol{y}_o}\{E_{\boldsymbol{\varphi}|\boldsymbol{y}_o,\mathcal{M}}\left[\boldsymbol{\varphi}\right]\}, \end{aligned} \tag{2.7}$$

with the first term representing within-model variance (due to the uncertainty encoded in the probability density function of uncertain model parameters and inputs $\boldsymbol{\Theta}_k$) and the second term representing between-model variance (conceptual uncertainty within the set $\mathcal{M}$ of considered models). Both terms result from applying the law of total variance with respect to the conceptual uncertainty within $\mathcal{M}$.

The posterior model weights $P\left(M_k|\boldsymbol{y}_o\right)$ reflect the probability of the individual models to be the most adequate one from the set in light of the observed data. The model weights are determined from Bayes' theorem:

$$P\left(M_k|\boldsymbol{y}_o\right) = \frac{p\left(\boldsymbol{y}_o|M_k\right) P\left(M_k\right)}{\sum_{i=1}^{N_m} p\left(\boldsymbol{y}_o|M_i\right) P\left(M_i\right)} = \frac{p\left(\boldsymbol{y}_o|M_k\right) P\left(M_k\right)}{E_{\mathcal{M}}\left[p\left(\boldsymbol{y}_o\right)\right]}, \tag{2.8}$$

with the prior belief $P\left(M_k\right)$ that model M_k could be the most adequate one in the set before the observed data have been considered. The denominator in Equation 2.8 normalizes the model weights such that they sum up to one. $p\left(\boldsymbol{y}_o|M_k\right)$ is referred to as Bayesian model evidence, marginal likelihood or prior predictive because it quantifies the average likelihood of the observed data based on a model's prior parameter and input space (*Drton et al.*, 2009):

$$p\left(\boldsymbol{y}_o|M_k\right) = \int_{\Omega_k} p\left(\boldsymbol{y}_o|M_k, \boldsymbol{\Theta}\right) p\left(\boldsymbol{\Theta}|M_k\right) d\boldsymbol{\Theta}_k = E_{\Theta|M_k}\left[p\left(\boldsymbol{y}_o\right)\right]. \tag{2.9}$$

$p\left(\boldsymbol{\Theta}|M_k\right)$ denotes the prior distribution of the model inputs and parameters, defined on the domain Ω_k. $p\left(\boldsymbol{y}_o|M_k, \boldsymbol{\Theta}\right)$ is the likelihood of a realization of model M_k to have generated the observed data set $\boldsymbol{y}_o$. The model evidence term can either be evaluated via integration over the whole input and parameter domain Ω_k (Equation 2.9) (*Kass and Raftery*, 1995), or via the posterior probability density function $p\left(\boldsymbol{\Theta}|M_k, \boldsymbol{y}_o\right)$ (Equation

2.2). Both alternatives pose a major technical challenge (see Section 2.5). The need to evaluate Bayesian model evidence for model ranking represents an evil difference to traditional applications of Bayesian updating. Obtaining posterior statistics from Equation 2.2 has become tractable only since the development of Markov chain Monte Carlo (MCMC) algorithms, because they entirely avoid the evaluation of model evidence by dropping the normalizing constant and evaluating only the proportionality.

The two levels of applying Bayes' theorem to obtain posterior model weights *and* individual posterior model predictions have been referred to as the "two levels of inference" by *MacKay* (1992). The updating step (Equation 2.2) and the model evaluation step (Equation 2.8) are obviously intertwined if the same data set $\boldsymbol{y}_o$ is used for both tasks. This fact means that the model comparison is based on uncalibrated models (or more generally, on models reflecting a state of knowledge prior to the analysis of the collected data $\boldsymbol{y}_o$) as done in this thesis, or, if this is not the intention of the modeler, a new data set needs to be collected to perform BMA as a validation step.

Interpretation of model evidence and model weights Note that model weights and model-averaged statistics are not only conditional on the calibration data set, but also on the chosen set of considered models. As such, model weights are expected to change if a new model enters the competition due to their joint normalization to sum up to one (see Equation 2.8). Bayesian model evidence itself, in contrast, is an objective likelihood measure independent of the set of models. Future model variants can be compared to the current set by using the same data set $\boldsymbol{y}_o$ and the same likelihood function, which makes model evidence "future-proof" (*Skilling*, 2006).

The ratio of model evidences for two competing models is referred to as Bayes factor (*Kass and Raftery*, 1995). It is equivalent to the ratio between the posterior and prior odds of two competing models M_k and M_l:

$$BF\left(M_k, M_l\right) = \frac{P\left(M_k|\boldsymbol{y}_o\right)}{P\left(M_l|\boldsymbol{y}_o\right)} \frac{P\left(M_l\right)}{P\left(M_k\right)} = \frac{p\left(\boldsymbol{y}_o|M_k\right)}{p\left(\boldsymbol{y}_o|M_l\right)}. \tag{2.10}$$

The Bayes factor measures the significance in the evidence of hypothesis M_k against the null-hypothesis M_l. Rules of thumb by *Jeffreys* (1961) and *Raftery* (1995) help the modeler to interpret Bayes factor values and to define threshold values at which a model should be dropped or selected from the set. However, even a model obtaining an almost diminishing model weight might still contribute significantly to between-model variance as found by *Rojas et al.* (2010) and *Wöhling et al.* (2015). Hence, model selection as

such can be questioned, while model averaging is more on the safe side when aiming for a comprehensive uncertainty assessment.

BMA can be understood as an extension of the Bayes factor to hypothesis testing for multiple models: based on model evidences, BMA allows to compare a hypothesis (a model) not only against a single null-hypothesis, but against a set of other hypotheses.

2.2 Definition of the Model Set

Representation of the model space A major challenge before starting any multi-model approach, including BMA, lies in the definition of the model set. Since all derived probabilities are conditional on this set, true probabilities could only be obtained if the model set was complete. Also, conceptual uncertainty in a strict sense could only be quantified if the model space was completely covered by the analysis. Since it is in reality both conceptually and computationally impossible to represent the entire model space, modelers need to put considerable effort into adequately sampling the model space and then into assigning realistic prior model weights.

Although the definition of the model set is widely acknowledged as one of the most critical issues in the context of BMA, there has been no real scientific progress even after several decades of research (*Jeffreys*, 1939; *Jaynes*, 1985; *Refsgaard et al.*, 2012). *Gull* (1988) already stated that "The real art is to choose an appropriate 'space of possibilities', and to date we have no systematic way of generating it." This statement still applies.

Distance between models Current research focuses on how the distance between models could be reasonably measured (e.g., *Abramowitz and Gupta*, 2008), because this allows a modeler to estimate how well the model space is currently sampled and how prior weights should be distributed to account for an over- or underrepresentation of a certain model type. The impact of prior model weights on the outcome of BMA results has been investigated by *Rojas et al.* (2009) and *Ye et al.* (2005), among others.

2.3 Definition of the Likelihood Function

Formal likelihood definition In order to implement Bayes' theorem, a likelihood function $p\left(\boldsymbol{y}_o|\boldsymbol{\Theta}\right)$ needs to be specified. The likelihood function reflects the modeler's trust in the accuracy and precision of the calibration data set $\boldsymbol{y}_o$. Thus, it represents the probability density function of measurement errors. Typically, for unbiased measurements, a Gaussian distribution centered about zero is assumed to describe random

measurement errors, with smaller deviations from zero being more probable than larger deviations. Correlations between measurement errors of different data points can be accounted for. Using a Gaussian likelihood function is also mathematically convenient for arriving at analytical solutions, e.g. in combination with a conjugate prior (*Box and Tiao*, 1973).

Accounting for model structural errors However, in real-world applications, the existence of normally distributed residuals (differences between model predictions and measured data) is often questioned. Non-Gaussian residuals can occur if measurement errors follow a non-Gaussian distribution, and/or if a prediction bias adds to the discrepancy between model predictions and observations. A bias in predictions is introduced by model-specific structural errors. These errors can be formally accounted for, e.g. with a statistical error model that is attached to the process-based simulation model. Such error models could also treat temporal autocorrelation in prediction bias, as typically encountered in hydrological models. A successful application of structural error models to improve hydrological predictions has been demonstrated by *Kuczera et al.* (2006), among others.

While acknowledging the existence of model errors and hence addressing the need for more than a single conceptual model is the main motivation for using BMA, the explicit statistical treatment of structural errors on top of the individual process-based models has so far not been attempted within the BMA framework. Especially in the context of BMA, but also for single-model Bayesian updating, it is important to remember that the likelihood function should only reflect assumptions about measurement noise. Model bias descriptions should instead be formulated as part of the model itself: model structural errors $\boldsymbol{e}$ need to be integrated into a model's parameter space $\boldsymbol{\Theta}$ as stated in Section 2.1. The reason for this is that a model's bias description needs to be updated (calibrated) just like a model's parameters in order to be able to apply the posterior bias description to future predictions. Any error that is incorporated into the likelihood definition, in contrast, cannot not be applied to forecasts, but only controls the degree of conditioning.

Informal likelihood approaches If, for a specific application, model structural errors are not explicitly accounted for, but turn out to be large compared to measurement errors, a strict likelihood definition based on the assumed distribution of measurement errors can lead to very low likelihood values over the whole parameter space of a model. Proper model calibration will then be almost impossible, and from a hypothesis testing

perspective, practically any model will be rejected. This issue has been addressed by informal likelihood approaches such as the generalized likelihood uncertainty estimation method (GLUE) (*Beven and Binley*, 1992): instead of specifying a formal probability distribution of measurement errors, model performance thresholds are defined that allow to identify "behavioral" parameter realizations. Such approaches represent a concession to situations where only poorly performing models are available.

2.4 Tradeoff between Performance and Parsimony

Differences in multi-model approaches Multi-model frameworks mainly differ in the way individual model predictions are obtained, and in the way model weights are assigned. In the case of BMA, a predictive distribution is obtained instead of a single most likely prediction. The uncertainty in predictions results from a specification of prior uncertainty in parameters, inputs, and calibration data. The ability to account for these sources of uncertainty is a major advantage of BMA over deterministic multi-model approaches that "forget" about these uncertainties once they have arrived at a calibrated model state.

Different options for model weights include assigning prior (e.g., equal) weights as typically done in climate change modeling (e.g., *Palmer and Raisanen*, 2002), optimized weights with the objective to maximize the likelihood of the combined prediction (e.g., *Raftery et al.*, 2005; *Ajami et al.*, 2007; *Wöhling and Vrugt*, 2008), or trade-off weights that balance a model's skill with some penalty for model complexity. BMA represents a special case of the latter, because it implicitly follows the principle of parsimony or Occam's razor (*Jefferys and Berger*, 1992; *Gull*, 1988) such that the BMA weights reflect a Bayesian trade-off between model performance and complexity. Note that the model averaging techniques with optimized weights cited above have been referred to as BMA in the literature. However, they do not use Bayes' theorem to obtain model weights, which is why I classify them as a distinct, alternative type of multi-model approach. In this thesis, the term BMA only refers to the corresponding framework that solves Bayes' theorem (Equation 2.8) to obtain model weights.

Measures for model complexity The fact that BMA can account for model complexity (or flexibility) is again related to its probabilistic setting. The flexibility of a model results from the interplay of the prior uncertainty in parameters with the prescribed model structure, resulting in a specific predictive distribution. Bayesian model evidence measures the level of agreement between the predictive distribution and the observed

data as an average likelihood. A precise and accurate distribution will obtain a higher model evidence than a wide predictive distribution. This mechanism of BMA is related to the bias-variance trade-off in statistics (see, e.g., *Burnham and Anderson*, 2003). If the variance in a model's predictions is large (low precision), a significant fraction of predictions will show a large bias which reduces the average likelihood; if the variance is low, chances are that the data are hardly covered by the predictive distribution (low accuracy) and again the average likelihood is low. Intuitively, an optimal compromise (trade-off) must exist between bias and variance in order to score a maximum model evidence value.

Note that the predictive variance of a model is not necessarily directly related to the size of the prior parameter space. While a direct relationship typically exists in statistical regression models, a physical model structure with a higher number of parameters may lead to an even narrower predictive distribution than a similar model with fewer degrees of freedom. This empirical observation reveals the difficulty of how to define model complexity: as briefly discussed in *Schöniger et al.* (2015a), one could quantify model complexity by (1) counting the number of adjustable model parameters (some information criteria rely on this concept, see Section 2.5), or (2) by performing a factor analysis to account for parameter correlations, or (3) by accounting for data-parameter sensitivity to assess a model's flexibility in prediction space instead of in parameter space. BMA builds on the latter concept and intuitively yields the most consistent measure of complexity for hydrosystem models out of these alternatives.

The tradeoff-mechanism in BMA I will explain the mechanism of the tradeoff between performance and complexity (or rather: parsimony) in BMA in the following. The prerequisite for a BMA analysis is a prior specification of model structures (see Section 2.2), parameter distributions and measurement error statistics (see Section 2.3). These subjective definitions are illustrated in Figure 2.1 for a simplistic case of three arbitrary models and a calibration data set containing two observations. Model M1 yields predictions $\boldsymbol{y}$ independent of its parameter value $\boldsymbol{\Theta}$ (i.e., the parameter is not sensitive to the data), model M2 shows a linear dependence between parameter values and predictions, and model M3 is nonlinear in its parameter.

The flexibility in parameter values needs to be specified in the form of probability density functions. For all three models, I have used two alternative specifications: a uniform prior $p1$ and a more informed Gaussian prior $p2$ (Figure 2.1b).

The likelihood function $p\left(\boldsymbol{y}_o|M_k, \boldsymbol{\Theta}\right)$ needs to be defined according to the assumptions

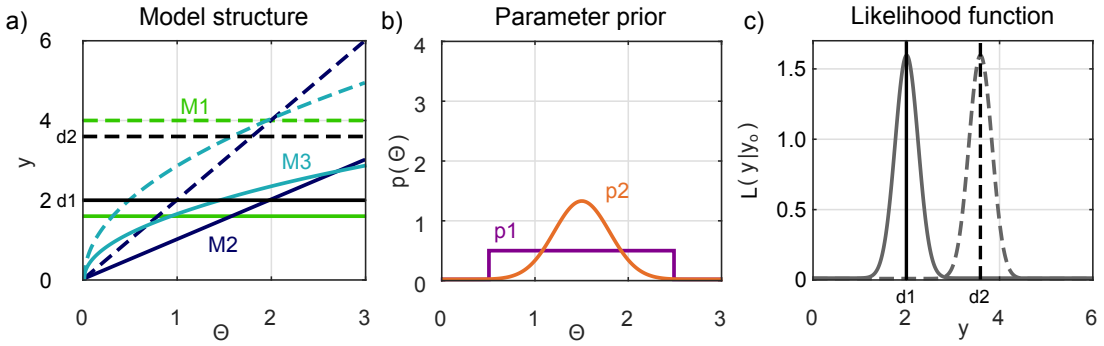


Figure 2.1: Subjective definitions required to perform BMA. a) Model set with predictions of data point $d1$ as solid lines and predictions of data point $d2$ as dashed lines, b) prior parameter probability density function with two alternative specifications $p1$ and $p2$, c) likelihood as a function of model predictions for data point $d1$ as solid line and data point $d2$ as dashed line.

about measurement noise. Here, a Gaussian distribution is assumed that is centered about an expected deviation of zero between model predictions and observed data (neglecting model structural errors for the sake of simplicity). Equivalently, one can understand this Gaussian distribution as a function of model predictions $\boldsymbol{y}$, centered about the observed data $\boldsymbol{y}_o$. Again, the highest likelihood is obtained for a mismatch of zero. I have chosen this interpretation for the illustration in Figure 2.1c, because it clarifies that the definition of the likelihood function is based on the characteristics of the observed data and that it is valid for all individual models (i.e., it should be defined independent of the models at hand).

As a function of the data $\boldsymbol{y}_o$, or predictions $\boldsymbol{y}$, or their mismatch $\boldsymbol{y} - \boldsymbol{y}_o$, the likelihood is a proper probability density function which integrates to unity. However, the likelihood can also be expressed as a function of model parameters, $L(\boldsymbol{\Theta}|\boldsymbol{y}_o)$ (*Fisher*, 1922). Assuming uncorrelated measurement errors, the likelihood of a parameter set to have generated both observed data values is determined as the product of the likelihoods per data point. The likelihood as a function of parameter values is displayed in Figure 2.2a. The likelihood function is now no longer a formal probability density function, since it is no longer normalized to integrate to unity. I choose this unnormalized likelihood as a function of parameters as the starting point for explaining what Bayesian model evidence values actually represent and how the tradeoff-mechanism works (Figure 2.2).

Based on a comparison of the likelihood functions, it can already be concluded that model M3 performs best out of the set presented here, because it achieves the highest likelihood among all models, and it achieves high likelihood values over a broad range

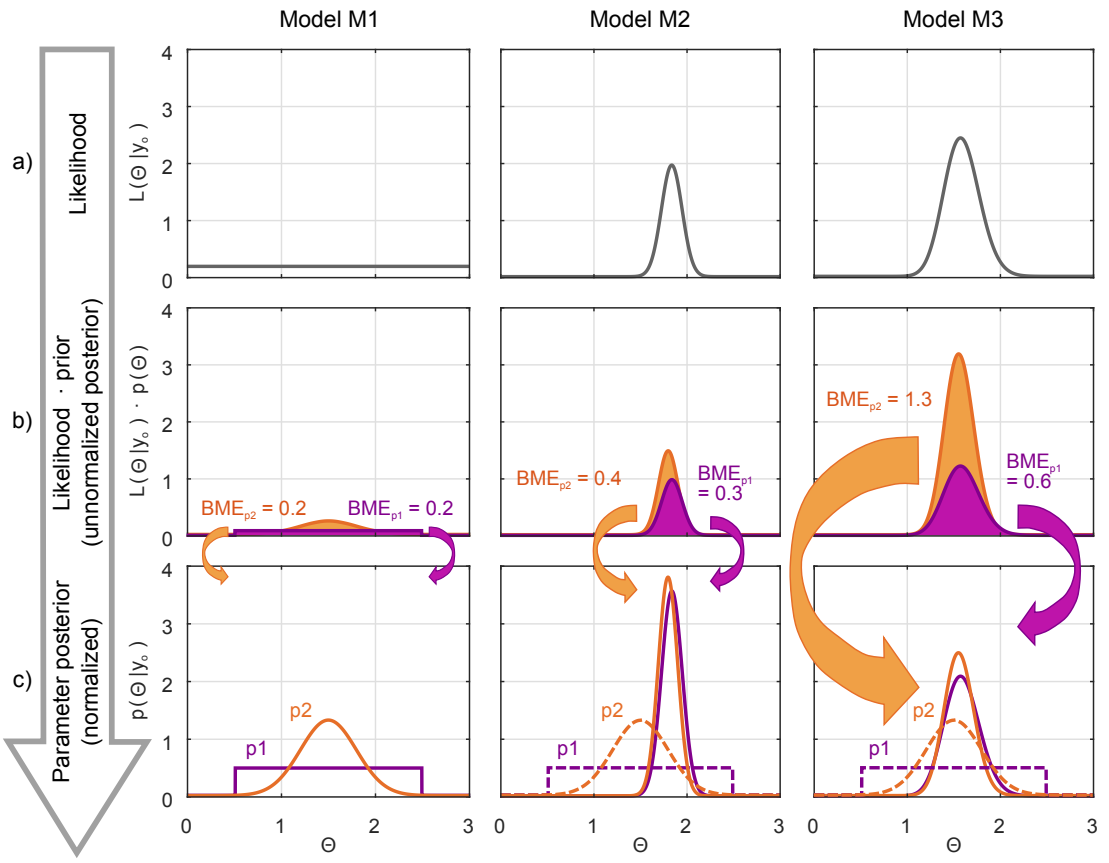


Figure 2.2: Objective BMA algorithm based on prior specifications (cf. Figure 2.1). a) Likelihood as a function of parameter values; b) likelihood multiplied with prior parameter probability, with Bayesian model evidence (BME) equal to the area under the curve; c) posterior parameter probability density function (solid lines) as compared to prior parameter probability density function (dashed lines).

of parameter values. This is an example where a nonlinear (intuitively more complex) structure yields a narrower predictive distribution: with varying parameter values, the nonlinear model still produces similar predictions that agree well with the observations. Hence, it achieves a high likelihood over a wider range of parameter values.

The differences in prior parameter probabilities now come into play: according to Bayes' theorem (Equation 2.2), the likelihood is multiplied by the parameter prior. This step can be understood as a reweighting of likelihood values. A wider parameter prior might have an adverse impact, because it needs to dilute its total probability mass of unity over a larger range of values. In the example of model M3, the uniform prior $p1$ causes a downweighting of likelihood values, whereas the sharper Gaussian parameter prior $p2$ causes a partial upweighting in those areas where high prior probability and high likelihood coincide (Figure 2.2b).

Finally, the product of likelihood and prior must be normalized to obtain the posterior probability density function of parameters (Figure 2.2c). The normalizing factor (cf. Equation 2.2) is exactly the targeted quantity Bayesian model evidence.

From Figure 2.2, two implications follow: (1) If parameters are sensitive to the data, the shape of the parameter prior influences the model evidence value (cf. models M2 and M3: for both models, the resulting model evidence value substantially differs between the two alternative prior specifications). (2) If parameters are not sensitive, BMA will not see a difference between alternative specifications of the parameter prior (cf. model M1). This type of (for this specific prediction target superfluous) model complexity is thus not tackled by BMA. Hence, the penalty for complexity is related to the calibratable parameter space of a model. A large reduction in uncertainty through conditioning will be penalized, leading to a lower model evidence value. Through this mechanism, BMA favors more robust models instead of "over-calibrated" models (see also the related discussion in *Schöniger et al.* (2014)). Thus, model complexity as seen through the eyes of BMA cannot be captured by measures of the prior parameter space nor of the posterior parameter space alone, but it is dependent on data-parameter sensitivity.

A typical practical situation might be that one model achieves a higher maximum likelihood, while the competing one achieves acceptable likelihood values over a wider range of parameter values. In that case, BMA will tend towards the more robust model, as long as the difference in performance is below the "tradeoff threshold". The ability to point the modeler towards an optimal balance between performance and complexity is a unique characteristic of BMA and distinguishes it from other multi-model frameworks. This trade-off must also be understood and considered when interpreting model ranking results: the reason for a model to win the competition could either be its superior performance, or its parsimonious structure, or a combination of both. Although the optimal trade-off is a core feature of BMA, its impact on model ranking results and the implications for future model improvement or model building strategies had not been systematically investigated prior to this thesis.

2.5 Technical Implementation

The drawback of BMA is that it requires the evaluation of Bayesian model evidence (Equation 2.9). For practical applications, typically no analytical solution to this integral exists, and numerical evaluation schemes require a high computational effort. Computing the posterior distribution of parameters instead (Equation 2.2) is similarly challenging.

Approximation of model evidence via information criteria Various authors have therefore proposed and applied approximations to the analytical BMA equations. *Neuman* (2003) introduced the Maximum Likelihood Bayesian Model Averaging approach (MLBMA), which relies on evaluating the Kashyap information criterion (KIC) at the most likely parameter set to avoid the integration over a model's parameter space. The Bayesian information criterion (BIC) or Schwarz' information criterion (*Schwarz*, 1978; *Raftery*, 1995) is a simplified version of the KIC that neglects prior knowledge. The Akaike information criterion (AIC) (*Akaike*, 1973) is derived from an information-theoretical background. Besides these very frequently applied information criteria, a long list of alternative versions have been proposed during the last decades (see, e.g., *Kadane and Lazar*, 2004).

Applicability of information criteria within BMA Many authors have, however, reported that these information criteria yield differing posterior model weights and even ambiguous model ranking results (*Poeter and Anderson*, 2005; *Ye et al.*, 2008, 2010; *Tsai and Li*, 2008; *Singh et al.*, 2010; *Morales-Casique et al.*, 2010; *Foglia et al.*, 2013). These findings suggest that the information criteria do not warrant the true Bayesian trade-off between performance and complexity, but instead point towards an arbitrary trade-off which is not in line with Bayesian theory. Prior to this thesis, *Lu et al.* (2011) made a first effort to investigate differences in the behavior of the KIC and the BIC as compared to a numerical reference solution in a synthetic geostatistical application. In general, there had been a lack of recommendations as to how BMA can be implemented efficiently and accurately.

2.6 Robustness of Model Weights

Sensitivity of model weights to calibration data Since Bayesian model evidence is a function of the chosen data set $\boldsymbol{y}_o$ through Equation 2.9, the outcome of model weights and of model ranking might change when using a different data set size or data type. The impact of the choice of calibration data on BMA results has become a recent focus of interest (*Rojas et al.*, 2010; *Lu et al.*, 2012; *Refsgaard et al.*, 2012; *Xue et al.*, 2014). I have contributed to this field of research in the course of a study on the worth of data for soil-plant model selection and prediction (*Wöhling et al.*, 2015). We have found that model ranking can vary substantially with the size and the composition of the data set.

This result is to be anticipated if the varied data sets "tell a different part of the story" about the underlying natural system. BMA can only judge to which degree the competing

models are able to predict the observed data (how fit the models are for this very purpose), but BMA cannot provide any more general insights regarding the degree to which the models actually represent the overall system to be modeled. Acknowledging that, through the eyes of BMA, the purpose changes with varied calibration data sets, a modeler needs to pay careful attention to the choice of the data set and to their judgment of whether the obtained BMA results are representative and robust.

Sensitivity of model weights to measurement noise Moreover, BMA weights not only depend on the chosen measurement design, but also on the very outcome of random measurement error for all individual data values. This functional dependence introduces uncertainty into the model weights. Yet, this source of uncertainty for model weights is not accounted for in the standard BMA routine. BMA weights are treated as fixed values, given a set of models and a specific data set. Therefore, prior to this thesis, there had been a need to acknowledge this additional source of uncertainty for model weights and to make its impact on model ranking results visible in an extended BMA routine.

Sensitivity of model weights to other sources of uncertainty Beyond measurement uncertainty in the calibration data set, other sources of uncertainty for model weights exist, such as noisy measurements of model forcings, or conceptual uncertainty in boundary conditions. The robustness of model weights against these sources of uncertainty needs to be assessed to find out whether the obtained model ranking is reliable or not.

3 Objectives & Contributions

This thesis aims to advance the available statistical tools for a consistent assessment of conceptual uncertainty in hydrosystem modeling. BMA is chosen as an integrated modeling framework because it can account for uncertainty in model input, in model parameters, in model structures, and in observed data. Resting on Bayesian probability theory, BMA is a statistically rigorous routine to obtain full predictive and parameter distributions which serve as a solid basis for decision making and scientific inference.

A wide-spread use of BMA in science and practice is, however, still hindered for several reasons. A significant extra effort is required to build a number of alternative models instead of a single one. Besides this obvious reason, applications of BMA are mostly hindered by difficulties with the technical implementation. Further, the interpretation of resulting model weights is non-trivial, given the implicit treatment of model complexity, data scarcity and different sources of uncertainty. To enable a confident and meaningful use of BMA, the following research questions are addressed in this thesis:

1. How can Bayesian model weights be evaluated efficiently and accurately?
2. How should model ranking results be interpreted in light of limited data?
3. How reliable are model ranking results under noisy input or calibration data?

Figure 3.1 places the three identified research questions into the context of modeling and uncertainty assessment of hydrosystems as schematically illustrated in Figure 1.1.

Part I: Evaluating Bayesian model weights The first part of this thesis addresses the technical challenge of how to evaluate BMA weights (see Section 4.1). In *Schöniger et al.* (2014), I have compared a comprehensive set of methods that can evaluate the required model evidence term with regard to underlying assumptions, accuracy, and computational effort. Five different mathematical approximations in the form of information criteria have been considered, as well as four numerical integration techniques. Based on a comparison of the underlying assumptions, I have argued which methods are truly suitable to evaluate model evidence, and which ones are prone to yield inaccurate results. I have further designed two synthetic test cases to investigate the impact of several influencing factors on the approximation quality in a controlled setup. Finally,

the evaluation methods have been tested on a real-world application of hydrological model selection. This systematic investigation of methods to determine Bayesian model evidence takes an important next step towards robust model selection, model averaging and uncertainty quantification in a BMA framework.

Part II: Assessing model justifiability in light of limited data If correctly evaluated, the main advantage of BMA as compared to other multi-model approaches is that it implicitly follows the principle of parsimony (Occam's razor) such that an optimal trade-off between goodness-of-fit during calibration and robustness of future predictions is identified. The second part of this thesis is devoted to the investigation of this trade-off. The objective is to answer the question, how much data are needed to reasonably calibrate a highly complex model, or, which level of complexity is justified given the available data? To this end, I have isolated the complexity component of the Bayesian trade-off from its performance counterpart. I refer to this analysis as *model justifiability analysis*, because it reveals whether any specific level of complexity can be justified by the available amount and type of data through the eyes of BMA. The proposed analysis can be run prior to the actual data collection in order to identify a most efficient measurement design. The model justifiability analysis has been introduced and demonstrated on a case of model selection between groundwater models of vastly different complexity in *Schöniger et al.* (2015a), see Section 4.2.

Part III: Accounting for sources of uncertainty for model weights The third part of this thesis extends the BMA framework to also account for uncertainty in model weights. With this new statistical concept, the robustness of model weights against uncertain calibration data or model input and the related confidence in model ranking can be assessed. I have proposed to investigate the variability in model weights with a resampling analysis (see Section 4.3). The BMA analysis is repeated for random outcomes of the uncertain variable, and the induced variability in model weights is analyzed. Such a resampling analysis reveals whether or not the competing models can be reliably ranked based on the chosen experimental data and the considered sources of uncertainty. In the specific case of weighting uncertainty due to noisy calibration data, a theoretical limit to model performance exists. The distance of individual models from this optimum can be assessed and interpreted to guide further model development. The proposed concept has been introduced and applied to soil-plant model selection in *Schöniger et al.* (2015b).

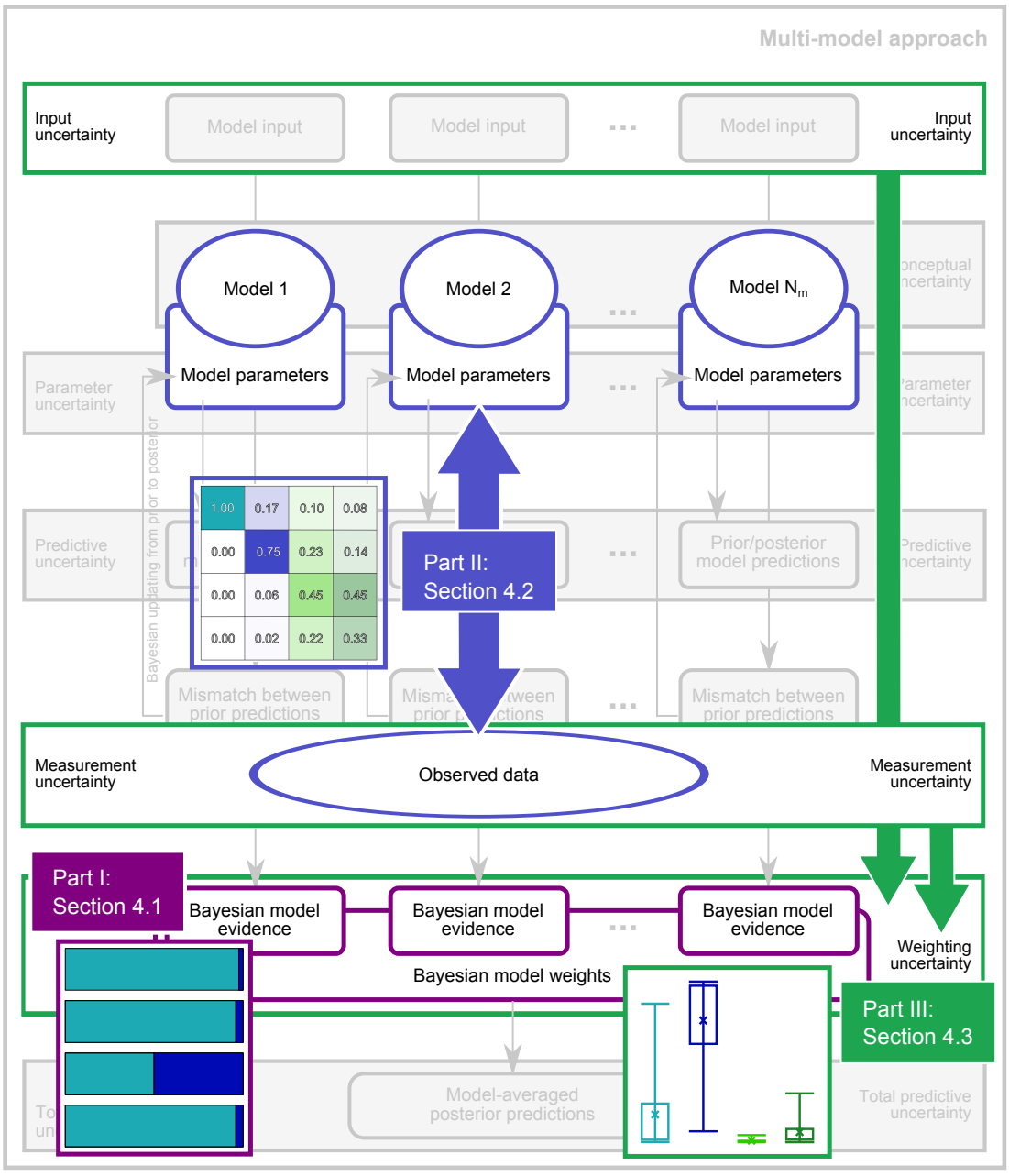


Figure 3.1: Research questions addressed by this thesis. (1) How to evaluate BMA weights? (2) How to choose a model in light of limited data? (3) How to assess the robustness of model weights against measurement noise and input uncertainty?

Expected impact The new statistical tools developed within this thesis are expected to be useful for a broad range of applications. These tools help to provide a more robust model ranking, an improved forecasting skill, a more accurate uncertainty quantification and hence better decision support for hydrosystem management questions, both in scientific research and in practice.

4 Results & Discussion

4.1 Evaluating Bayesian Model Weights

The evaluation of BMA weights poses a major computational challenge as pointed out in Section 2.5. I have addressed the challenge of determining the model evidence term (Equation 2.9) by comparing a set of different evaluation methods in *Schöniger et al.* (2014). Three classes of methods are in principle available: (1) analytical solutions which rarely exist for real-world applications, (2) computationally efficient mathematical approximations in the form of information criteria and (3) computationally demanding, but typically much more accurate numerical integration schemes. For the comparison, I have selected five information criteria that BMA is commonly operated with (see Section 2.5): the AIC, the AICc (bias-corrected AIC), the BIC, and two versions of the KIC. As representatives of numerical evaluation schemes, I have considered three classical Monte Carlo integration techniques (*Hammersley and Handscomb*, 2013): simple Monte Carlo integration, Monte Carlo integration with importance sampling, and Monte Carlo integration with posterior sampling. As a fourth numerical option, I have chosen a very recent approach called nested sampling (*Skilling*, 2006; *Elsheikh et al.*, 2013). I have first compared the underlying assumptions of these two classes of techniques, based on a comprehensive literature review.

Information criteria The BIC and the KIC are a result of the Laplace approximation to solve the integral in Equation 2.9, under the assumption of a Gaussian posterior parameter distribution. This assumption is fulfilled, e.g., if both the parameter prior and the likelihood function are Gaussian distributions and if the model is linear. In that case, the KIC evaluated at the maximum a posteriori parameter estimate (referred to as KIC@MAP) yields the exact analytical solution. In contrast, the KIC evaluated at the maximum likelihood parameter estimate (referred to as KIC@MLE) and the BIC deviate from that exact formulation and are therefore expected to yield an inferior approximation quality. The AIC and the AICc are based on an information-theoretical measure and were not specifically developed as approximations to Bayesian model evidence. Hence, they are not expected to provide an exact solution, but still, they are widely used in the context of BMA. The approximation error of all information criteria considered here depends on the actual application.

Numerical evaluation methods The accuracy of numerical methods, in contrast, is typically only limited by the affordable number of samples, i.e., by restrictions in computational effort. The challenge with regard to numerical evaluation methods is thus to find a scheme that yields a reasonably accurate estimate of model evidence with a reasonably small sample size. Out of the listed numerical schemes, simple (brute-force) Monte Carlo integration is the favored approach, because it neither introduces any bias into the model evidence estimate, nor requires any assumptions on the involved distributions or model structures. Yet, simple Monte Carlo integration is also the most computationally demanding numerical method considered here. Importance sampling and (even more so) posterior sampling reduce the computational effort, but at the same time introduce a bias into the model evidence estimate. In line with the folkloric no-free-lunch-theorem of optimization, there is a tradeoff between accuracy and efficiency, not only between the cheap information criteria and the expensive numerical methods, but also within the set of numerical schemes.

Nested sampling is a promising new integration method that transfers the integration from the high-dimensional parameter space to a one-dimensional likelihood space. Since the transfer rule is not perfectly known, some uncertainty arises from this approach and more research is needed to reduce this uncertainty; however, in principle, this and related approaches may represent more efficient alternatives to traditional integration schemes that require a high number of samples for convergence in high-dimensional applications.

Computational effort Regarding the computational effort of the different methods to evaluate model evidence, it needs to be considered that the computational efficiency of the information criteria varies with their formulation: while the KIC@MAP is the most favorable one among the criteria investigated here, it is also the most computationally expensive one. Its computational effort is comparable to numerical integration with posterior sampling, e.g. through an MCMC algorithm. Nevertheless, obtaining a posterior sample from MCMC is still much cheaper than performing simple Monte Carlo integration. Hence, although the KIC@MAP might not be a typical representative of a "cheap approximation", it potentially offers the chance of using MCMC results to perform BMA. This would be a major breakthrough: while performing Bayesian updating for individual models with the help of MCMC methods has become increasingly popular, its output can so far not be used to perform BMA. Be referred to *Schöniger et al.* (2014) for a detailed explanation why using MCMC output in posterior sampling yields a biased approximation of model evidence. If MCMC results could be used for BMA via the KIC@MAP, this would certainly trigger a more wide-spread use of BMA.

Accuracy in linear and nonlinear synthetic test cases To test the actual performance of the nine different model evidence evaluation techniques considered here, I have first designed two synthetic test cases with the goal to investigate the resulting approximation errors in a controlled setup. From the review of the theoretical background of the information criteria, I have identified three main factors that influence their approximation quality: (1) the size of the calibration data set, (2) the shape of the parameter prior, and (3) the non-linearity of the model. The size of the data set determines the relative importance of goodness-of-fit in the tradeoff with parsimony. As explained in Section 2.4, the shape and the dimensionality of the prior parameter distribution contribute to the flexibility (complexity) of a model. And finally, the degree of non-linearity in a model determines to which degree the assumption of the Laplace approximation is violated.

These three factors also influence the efficiency of the numerical schemes, but the accuracy achieved by the numerical schemes can be increased by increasing the number of samples. This is a major advantage over information criteria, because the convergence in the model evidence estimate with an increasing amount of computational effort provides some indication about the error of the approximation. The fixed result obtained from the information criteria, in contrast, does not allow any conclusion about their distance from the true solution.

To be able to determine the usually unknown approximation error of model evidence evaluation, I have in a first step set up a simplistic linear test case such that an analytical solution for the integral in Equation 2.9 exists. With the analytical solution at hand, I can compare the results of all nine model evidence evaluation methods with the true value and I can determine the respective absolute and relative errors. Further, I can assess the deviation from the true model ranking, based on the evidence values of competing models. This allows a first-time benchmarking of the different information criteria and the numerical evaluation schemes against the true solution.

With the linear test case, the influence of the two factors (1) data set size and (2) shape of the prior can be systematically investigated. This setting represents an ideal premise for the KIC@MAP and its simplified variants KIC@MLE and BIC, because their underlying assumption of a Gaussian parameter posterior is fulfilled. For this best-case scenario, the approximation error of these information criteria is expected to be the lowest possible.

The benchmarking of the different methods against the exact analytical solution has shown that the AIC(c) and the BIC yield arbitrarily large errors when approximating Bayesian model evidence. Unfortunately, their errors depend on the actual data set

used to perform BMA, and even on the outcome of measurement error in this data set. Even worse, these criteria cannot decipher prior information about the parameter or prediction space. This is alarming, because the prior information encodes the flexibility or complexity of a model (see Section 2.4). It is often argued that the AIC(c) or BIC would yield acceptable model evidence estimates if prior information is vague or not available. In this test case, though, I have demonstrated that the approximation quality even deteriorates for less and less informative priors.

The qualitative behavior of approximation error as a function of the two influencing factors is shown in Figure 4.1a and b. The outcome of model ranking for two arbitrarily chosen linear model structures (light and dark blue bars, respectively) is illustrated in Figure 4.1c. Note that the error made by the KIC@MAP is not displayed because it is zero in this linear case.

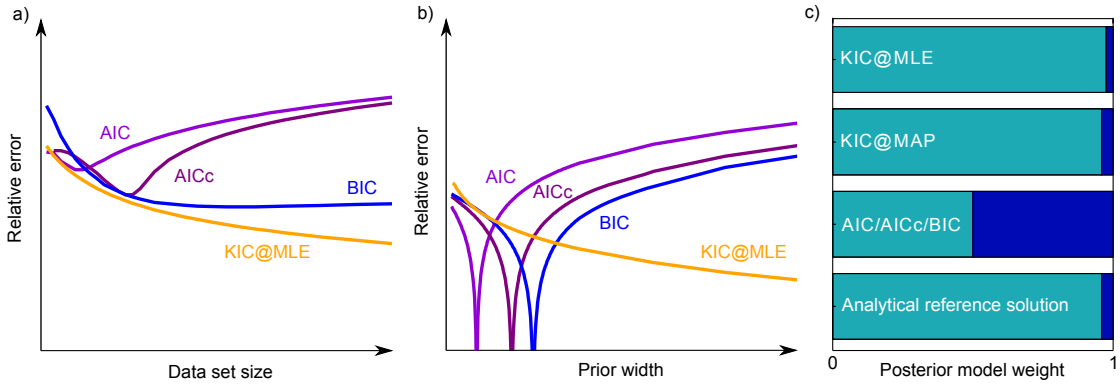


Figure 4.1: Qualitative illustration of errors made when approximating Bayesian model evidence by information criteria for the simplest case of linear models (modified from *Schöniger et al.* (2014)). Behavior of relative error in evidence approximation as a function of (a) data set size and (b) width of the parameter prior. (c) Corresponding deviations from the true model ranking.

To investigate factor (3), the non-linearity of the model, I have in a second step created a slightly more challenging synthetic test case that involves nonlinear models. In this case, an analytical solution does not exist anymore, and the assumptions behind the KIC and the BIC are no longer fulfilled. To be able to judge the quality of approximation, I use the result of simple Monte Carlo integration with a very large number of samples as reference, because this method has proven to be the most accurate one both in the theoretical review and in the linear test case. While the KIC@MAP provides the correct solution for Bayesian model evidence when its assumptions are fulfilled (e.g., in the linear test case), its approximation quality deteriorates rapidly when confronted with nonlinear models.

As explained before, the quality of approximation by the numerical integration schemes is a function of computational effort. Results from the synthetic test case have confirmed the expectation that simple Monte Carlo integration and Monte Carlo integration with importance sampling achieve the highest accuracy and lowest numerical uncertainty at a specific computational effort. Nested sampling produces slightly higher approximation errors, but still outperforms all tested information criteria on average over a large number of measurement error realizations (be reminded that the error made by the information criteria varies substantially with the chosen data set). Nested sampling further yields a very accurate model ranking, which indicates that its approximation error is consistent over the competing models.

Accuracy in real-world test case In a third step, the nine model evidence evaluation methods have been tested on a real-world application of hydrological model selection. The discharge of the Fils river in the Upper Neckar basin in Southwest Germany is predicted by two versions of the distributed mesoscale hydrologic model (mHM) (*Samaniego et al.*, 2010). The two models differ slightly in their conceptualization of soil structure. Model ranking is performed based on a nine-year time series of daily discharge observations. Due to the non-linearity in these two models, again no analytical solution for Bayesian model evidence exists, and the assumptions behind the KIC and the BIC are violated. As reference solution, I have again performed simple Monte Carlo integration with a vast amount of realizations to ensure convergence of the model evidence estimate.

Results of this real-world test case have confirmed that the KIC achieves the best approximation quality out of the information criteria considered here. However, it still performs much worse than any of the numerical integration schemes. Model ranking based on this large data set yields a very clear proposition for model choice according to the reference solution. Despite the significant errors in approximating model evidence made by the information criteria, this clear ranking is surprisingly reflected by all of the considered evaluation methods. One exception is the BIC, which yields a completely reversed model ranking. The inferior performance of the BIC came without warning, since the BIC did not perform particularly poorly in the two synthetic test cases. The potential for such unexpected behavior must, on the other hand, be anticipated when recalling my previous findings that the performance of the AIC(c) and the BIC depends on the application and the data set at hand and that the model ranking suggested by those information criteria is somewhat arbitrary and not necessarily correlated with the true one.

The unambiguous model ranking in this case results from the fact that, if a sufficiently large data set is available, one of the models will receive a weight of close to 100 %, no matter how minor the differences in model structures might seem. Several authors had suspected a causal relationship between the use of information criteria and such very decisive model ranking results, and proposed specific modifications to correct this behavior (e.g., *Tsai*, 2010; *Lu et al.*, 2013). Based on my investigations, this behavior is *not* an artifact of using information criteria, but a characteristic of BMA, or more specifically of the definition of the likelihood function (in this case, a specific characteristic of assuming uncorrelated errors). When applying BMA, a modeler must carefully choose the type and extent of calibration data and the type of likelihood function. Future studies on alternative options to define the likelihood function should always be guided by a numerical reference solution for the BMA weights, instead of relying on information criteria that are prone to yield a biased result.

Ranking the same two hydrological models based on a much shorter time series of observations has shown that, if the goodness-of-fit term is not as dominant over the parsimony term, model ranking is slightly more ambiguous, and only the numerical methods are able to reproduce the true result with satisfying accuracy.

Summary and implications Based on the performance of the considered information criteria in the three test cases, I conclude that the AIC and the BIC are not able to provide a reliable approximation to Bayesian model evidence. Further, the errors are not consistent in the sense that they would be compensated when calculating the ratio of evidence values for model ranking. Even for linear models, model ranking could not be satisfyingly reproduced by the AIC and the BIC, although the test case settings represent a best-case scenario regarding their performance. Any more successful application in a real-world test case would be pure luck. These information criteria are not able to detect the true dimensionality of a model, and as a consequence, they yield trade-off weights that generally do not reflect Bayes' theorem. Since it cannot be foretold how big the price (the lost accuracy) will be that one has to pay for saving computational effort, I do not recommend to use the AIC or the BIC in BMA. While the KIC@MAP provides the exact solution when its assumptions are fulfilled, its performance deteriorates rapidly if they are violated, which is unfortunately typically the case in real-world applications. Then, only numerical integration can provide a reliable approximation to the true Bayesian model ranking. Be referred to *Schöniger et al.* (2014) for an overview table of the merits, pitfalls and recommended uses of the investigated evidence evaluation techniques.

4.2 Assessing Model Justifiability in Light of Limited Data

The characteristics of the BMA trade-off between performance and complexity are the focus of the second part of my thesis. This trade-off is needed if only few data are available. With few data points, it is very difficult to distinguish with confidence the true or best performing model from competitors. This is very similar to the calibration process of a model: a complex model needs much more data to be well-defined in its parameter values than a simple model. The philosophy of BMA is "when in doubt, prefer the simpler model". Let us assume for the sake of the argument that the true model is actually in the chosen set of models (which will never be the case in reality). Then, BMA will still prefer simpler models than the true model up to a certain amount of informative data, because the true complexity could not yet be justified by the available data. But with an infinite number of data points, BMA will identify the true model with perfect confidence, no matter how simple or complex it might be. This behavior is an important trait of a model selection framework, but it is not guaranteed by other model ranking techniques. Model selection with the information criteria AIC or BIC, e.g., tend to over- or underestimate the true dimensionality, respectively (*Burnham and Anderson*, 2004).

The concept of model justifiability I refer to the successful identification of the true model through BMA as *model self-identification*. Simple true models can be self-identified with small data sets, while complex true models need much larger data sets to be justified. To test whether the maximum degree of complexity that can still be self-identified through BMA is actually rising with increasing data set size and falling with decreasing data set size, I have developed an analysis tool in a synthetic setting. A set of models is chosen that covers a wide range of complexity. Then, each of the models is used to generate a number of synthetic data sets, i.e., samples from the predictive distribution of a model are now used as data sets. Based on each of these data sets, the standard BMA analysis is performed to obtain model weights. Now, these model weights are averaged over all those data sets (predictions) that were generated from one specific model. In the end, one obtains as many sets of model weights as the number of compared models N_m. These average weights then build an $N_m \times N_m$ matrix which I call *model confusion matrix*. Confusion matrices are known from the field of machine learning. Such a matrix summarizes the performance of a classifier algorithm by distinguishing between correctly classified and wrongly classified (confused) items.

I have transferred the concept of a confusion matrix to the context of model selection. The model confusion matrix indicates the highest degree of complexity that could possibly be justified (self-identified) based on the chosen experimental setup. The complexity of a model is deemed justifiable if it receives the highest model weight in the set when it has generated the data, i.e., when this model's predictions have been used as calibration data sets. A model weight of close to one means that the model's complexity is perfectly justifiable, whereas almost uniform weights indicate a highly uncertain justifiability. The maximum justifiable degree of complexity is the complexity of the most complex model in the set which still achieves justifiability. Figure 4.2a shows an arbitrary example of a model confusion matrix. In this case, the degree of complexity inherent to model M3 would be identified as the maximum justifiable degree given the chosen experimental design. Figure 4.2b schematically illustrates how curves of model weights for the data-generating model (i.e., the average weights on the main diagonal of the model confusion matrix) depend on model complexity and data set size.

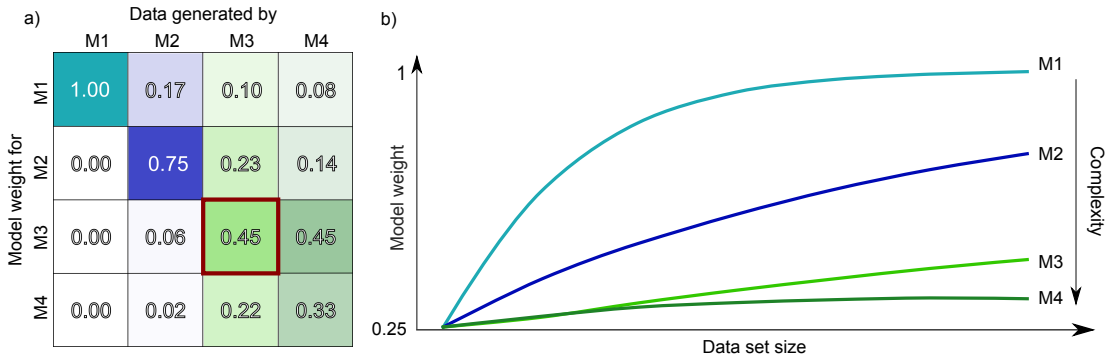


Figure 4.2: Model justifiability analysis, exemplarily shown for a set of four models M1 to M4 (modified from *Schöniger et al.* (2015a)). (a) Model confusion matrix, with the maximum justifiable degree of complexity identified as the complexity of model M3 (marked in red). (b) Model justifiability (average model weight for the data-generating model) as a function of complexity and data set size.

The model confusion matrix provides two major benefits: (1) It reveals the degree of similarity between the alternative models. This knowledge helps a modeler to reconsider the choice of prior model probabilities, because one might want to dilute prior model weights if some models appear to share a specific structure (*George*, 2010). (2) It helps to interpret the model weights resulting from the conventional BMA procedure. If two or more models receive a very similar weight, it is typically difficult to decipher whether the models are actually very similar in their predictions, or whether the similar model weights just result from a similar overall goodness-of-fit. The model confusion matrix can clearly discover the former case.

Application to groundwater model selection The justifiability analysis is proposed in *Schöniger et al.* (2015a). To illustrate the analysis, I have chosen an application to aquifer characterization via hydraulic tomography. Steady-state drawdown data are used to characterize the heterogeneous spatial distribution of hydraulic conductivity in a sandbox aquifer. The construction of the sandbox and the experimental design for hydraulic tomography are documented in *Illman et al.* (2010). Four alternative concepts to parametrize the spatial distribution of hydraulic conductivity are considered: (1) a homogeneous, equivalent medium approach, (2) a geological model that is inspired by the true layering visible from the sandbox, (3) a geostatistical interpolation by pilot points, and (4) a full geostatistical approach. These four approaches differ vastly in complexity and computational effort. Further, the required amount of calibration data clearly increases with model complexity. Geostatistical approaches have proven to yield skillful and reliable prognoses also beyond the calibration setup when calibrated with a comprehensive data set from hydraulic tomography (*Illman et al.*, 2010). On the other hand, the much simpler geological model with zones of constant hydraulic conductivity is more frequently applied in practice and in science, and has also shown acceptable results in comparison to geostatistical approaches (*Berg and Illman*, 2011; *Illman et al.*, 2015). Hence, the question arises whether, under a limited amount of data, the highly flexible geostatistical approach is justified, or whether the more robust zonated model should be preferred from a parsimonious point of view. This question is answered by applying BMA to rank these four different parameterizations.

I have used this test case to find out how much data are required to justify the different proposed levels of complexity, ranging from one effective parameter (homogeneous approach) to a full geostatistical approach where hydraulic conductivity varies in each of the 12,480 grid cells used for spatial discretization. I have further investigated how the maximum level of complexity that can still be justified depends on the amount of available data. To this end, I have used hydraulic tomography data from between one and six pumping tests for the inversion and repeated my suggested justifiability analysis for all these data sets.

Results have shown that the simplest model (homogeneous, equivalent medium) can be (almost) perfectly justified in all data set variants, even if only a single pumping test is used for hydraulic tomography. Justifiability of the more complex geological model can also be achieved in all data set variants, however, with less confidence: when the geological model generates the data sets, it does obtain a higher model weight than the competitors, but its weight does not (yet) approach 100 %. Justifiability becomes even

more difficult in the case of the interpolation model. Its complexity cannot be supported by a single pumping test anymore and, when adding more data, it still does not obtain a model weight of more than 50 %, i.e., it never obtains an "absolute majority". The most complex full geostatistical model is the only one which cannot be justified, not even by the most comprehensive data set considered in this study. When this model generates the data, the relatively similar interpolated model obtains the largest weight out of all models. Thus, the maximum level of complexity that can be justified through the eyes of BMA with the data set variants investigated here is represented by the complexity of the interpolated model. I suspect that many more informative data points would be needed to reach a breaking point such that the full geostatistical model could be justified as well.

I have further observed that the more data are used for the BMA ranking, the higher the confidence of justifiability. This behavior agrees with the theoretical claim that BMA will recognize the true model, regardless of its complexity, in the limit of an infinite data set size. Also, the test case results have confirmed that BMA will consistently point towards the true model even if it is of low complexity and a large data set is used. Hence, BMA does not tend to over-fit the data as opposed to, e.g., model ranking with the AIC (*Burnham and Anderson*, 2004).

Interpretation of model ranking results based on justifiability After having performed my proposed justifiability analysis in a synthetic setup first (the experimental setup has been used in the analysis, but only synthetic, model-generated data instead of observed data), I have performed the standard BMA analysis in a second step. While the first step indicates which level of complexity could possibly be retrieved from the chosen experimental setup, the second step shows how the models are ranked based on the actually observed experimental data. Then, these results are interpreted with the help of the findings from the first step.

When using the actually observed drawdown data for BMA, the geological model always ranks first, no matter which data set variant (observations from one to six pumping tests). As explained in Section 2.4, there are three potential explanations why this specific model scores best: either, it is parsimonious enough in the eyes of BMA, or it shows a superior goodness-of-fit, or because of a mixture of both reasons. To disentangle these potential reasons, I have drawn upon my findings from the justifiability analysis in the first step. Since the geological model can be well self-identified under the current experimental design, its high model weight when confronted with the observed data suggests a close agreement between the model's predictions and the observed system response.

To cancel out the other option that the geological model has only been ranked first due to its simplicity, I have set up another competing model of almost the same complexity. The competing model structure completely ignores the knowledge about the packing pattern of the sandbox. I have repeated the BMA analysis with the original geological model replaced by this "impaired" version of it. The geology-ignoring variant scores significantly worse and is ranked last in all data set variants. Thus, simplicity is not the key to a high model weight in this case, and I have concluded that the geological model with the informed structure wins the BMA context because of a reasonable amount of flexibility at the right spots.

For this specific test case application, I have found that the inversion of hydraulic tomography data does not per se justify or require a geostatistical description of aquifer heterogeneity. Rather, a well-informed geological model might provide more robust results. Combining these two approaches by equipping a zonation-based model with a geostatistical representation of small-scale variability in hydraulic conductivity within zones (see, e.g., *Fienen et al.*, 2008) seems a promising approach that should be further investigated in science and finally applied in practice.

Summary and implications I have performed a two-step BMA procedure to find out (1) which groundwater model complexity could possibly be justified by using data from one to six pumping tests for hydraulic tomography, and (2) which model actually turns out to incorporate the optimal Bayesian trade-off between performance and complexity. The proposed two-step procedure disentangles model justifiability (in light of the experimental setup) from model adequacy (in light of the actually observed data). Since model justifiability is assessed in a synthetic setting, it can be used to guide the planning of experiments. This analysis will tell modelers which data to collect in order to run a meaningful model ranking analysis. Modelers should be aware that, if the available data are not sufficient to justify the most complex model under consideration, the most complex model could still be the one closest to the true underlying system, but it will not be ranked first by BMA, due to the implicit preference of parsimonious models.

The additional computational effort for the proposed justifiability analysis is relatively low compared to the effort for the numerical evaluation of the conventional BMA algorithm, which is the most reliable method of evaluation (see results of the first part of my thesis, Section 4.1). Therefore, I recommend to perform both steps in applications for which the interpretation of resulting BMA weights might seem non-trivial. I expect this to be the case in the vast majority of real-world applications.

4.3 Accounting for Sources of Uncertainty for Model Weights

The third part of this thesis focuses on the reliability of model weights under different sources of uncertainty. While, traditionally, model weights are perceived as fixed values, I have claimed that they are sensitive to uncertainty in the calibration data set. Traditionally, the BMA framework is able to account for the impact of this source of uncertainty on the predictive distribution of individual models, and assigns fixed model weights based on the agreement between the predictive distributions and the observed data. This routine neglects that the weighting would turn out differently if the measurement error in the observed data set had taken on a different value (e.g, if measurements were repeated). Hence, I have proposed to treat model weights as uncertain quantities as well (*Schöniger et al.*, 2015b).

The need for an extended BMA routine Accounting for measurement noise as source of uncertainty for model weights is a logical consequence and improvement of the standard BMA scheme, because assumptions about the statistical properties of measurement noise need to be specified anyway in order to define the likelihood function. Given these assumptions, a modeler should test whether the obtained weights are robust under varying outcomes of measurement error in the data set. As soon as an assumption about the quality of the data is made, my proposed routine should follow.

Other sources of uncertainty for model weights such as uncertain forcings or boundary conditions could be treated with my suggested extension in the same fashion, although it might seem less natural because assumptions on the quality of model input need to be made, which are not required (but may be included) in the standard BMA routine.

The statistical concept of resampling I have proposed to assess the variability of model weights under any specific source of uncertainty by a resampling analysis within a Monte Carlo framework. The uncertainty in an uncertain quantity ω (e.g., measurement error in input or output data) is represented by generating a sufficiently large number of random realizations from its assumed distribution $p(\omega)$. This procedure is referred to as parametric bootstrap (e.g., *Davison and Hinkley*, 1997). For each of the random realizations and each of the models, the corresponding model evidence is evaluated. The outcome of the resampling analysis is then an ensemble of evidence values and model weights.

The variability in model weights can be summarized by determining confidence intervals (exemplarily shown in Figure 4.3a). By assessing the overlap between confidence intervals of model weights for competing models, first insights about the robustness of model ranking under the investigated source of uncertainty can be gained. The distributions of model evidence allow to assess the reliability of model ranking in further detail (Figure 4.3b). Several options are available to analyze and interpret the resulting distributions of model evidence as discussed in *Schöniger et al.* (2015b). As most promising, I have identified the Bayes factor (Equation 2.10). From the distributions of Bayes factors, the modeler can tell how reliable a statement like "model A scores a higher evidence value than model B" actually is. Such a detailed analysis is of great worth for further model development.

Theoretical limit to model performance In the specific case of considering measurement noise in the calibration data as source of uncertainty for model weights, statistics of model evidence values can be used to find out how far off the models' performances are from the theoretical maximum. This theoretical limit for model performance exists, because even a perfect (true) model would yield a distribution of model evidence values under noisy calibration data. I define the *theoretically optimal model* (TOM) as the observed data set, because it combines a perfect fit with zero complexity (no adjustable parameters). The TOM's evidence distribution represents the best-case scenario of a model evidence distribution under noisy data. The distance of the competing models from the TOM or their overlap, respectively, can then be measured in the same fashion as the evidence distributions of competing models are compared with each other, e.g. with the help of Bayes factors (*Schöniger et al.*, 2015b).

In the spirit of strict hypothesis testing, a modeler could choose a significance level and check whether the proposed model fails this test against the TOM with a specific level of evidence (e.g. using the rules of thumb mentioned in Section 2.1). This is schematically illustrated in Figure 4.3c.

The comparison of a model's performance with the theoretically optimal performance (the performance of the TOM) provides an absolute measure for model skill, as opposed to relative model weights that are conditional on the current model set. After having evaluated this absolute distance, a modeler is better able to decide whether the already considered models perform sufficiently well for the given purpose, or whether they need to be improved, or whether completely new models should be included in the comparison.

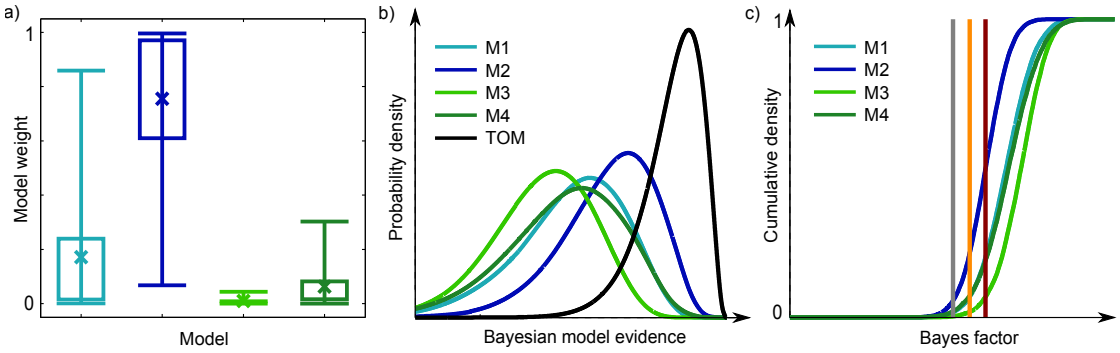


Figure 4.3: Robustness of model ranking against measurement noise in the calibration data set, exemplarily shown for a set of four models M1 to M4 (modified from *Schöniger et al.* (2015b)). (a) Boxplots showing mean values and confidence intervals of possible model weight outcomes, (b) distribution of possible model evidence values, (c) distribution of Bayes factors in favor of the theoretically optimal model (TOM), with colored lines indicating specific significance thresholds.

As a welcome side effect, the proposed analysis tools are expected to trigger further research on the question how to sample the model space well (see Section 2.2). Knowing some sort of distance between competing models and between the individual models and the TOM is a good basis to start from, because the model space can only be extended efficiently with new or modified models if the current sampling state is well characterized.

Application to soil-plant model selection I have demonstrated the use of the extended BMA routine in an application to soil-plant modeling, following up on a previous study I was involved in during the course of my PhD project (*Wöhling et al.*, 2015). In the cited study, we have performed the standard BMA analysis for four crop models that differ in how detailed and how mechanistically plant processes are represented, but are coupled to the same model to simulate water movement through the soil. The BMA analysis has greatly helped to diagnose model structural deficiencies of the four alternative crop models, and to understand how they react to different calibration data types. We have further investigated, how the amount of data influences the decisiveness in model ranking.

I have continued along this path by investigating how model weights change under random outcomes of measurement noise in the calibration data set, and how model ranking results are affected (*Schöniger et al.*, 2015b). Assumptions about the measurement error distributions were derived from replicated measurements and experiences from previous experiments.

To perform the proposed resampling analysis, I have drawn random realizations of measurement error from its specified distribution and I have added them to the observed data values since, on average, measurement error is assumed to be zero. Model evidence values are determined from simple Monte Carlo integration which has been proven to yield most accurate results in the first part of my thesis (see Section 4.1). Distributions of model evidence are then obtained by repeating the standard BMA analysis for random outcomes of measurement error.

I have repeated the full resampling analysis for different calibration data sets. Considering different data types for calibration allows to draw conclusions about the relevance of weighting uncertainty under different calibration conditions. I further went through different assumptions on the shape of the distribution of measurement error to find out how an increase or decrease in noise influences the robustness of model ranking.

Reliability of model ranking Results of this study have shown that the variability in model weights due to noisy data is especially high for the two best-performing models in this specific application. This variability leads to ambiguous model rankings in all investigated calibration scenarios. It can be concluded that, based on the given experimental design and the specified assumptions about measurement noise, no single model can be selected from the set with confidence.

A reduced measurement error standard deviation has yielded more decisive results, while an increased measurement error standard deviation has led to even more variability. Hence, to arrive at a more solid model ranking, more accurate data or other more informative data types should be collected. The search for most useful future data could be guided by optimal design approaches that rest on the proposed extended BMA routine.

Propagation of weighting uncertainty to total predictive uncertainty Beyond the impact on model ranking results, weighting uncertainty intuitively must also have an effect on the total uncertainty of model-averaged predictions. Therefore, the total variance formulation of the standard BMA routine needs to be extended to account for weighting uncertainty. I propose to derive the extended variance formulation from Equation 2.7 using the law of total variance:

$$
\begin{aligned}
V_{\varphi|y_o}\left[\varphi\right] = \quad & E_\omega\left\{V_{\varphi|y_o,\omega}\left[\varphi\right]\right\} + V_\omega\left\{E_{\varphi|y_o,\omega}\left[\varphi\right]\right\} \\
= \quad & E_\omega\left\{E_{\mathcal{M}|y_o,\omega}\left[V_{\varphi|y_o,\mathcal{M},\omega}\left[\varphi\right]\right] + V_{\mathcal{M}|y_o,\omega}\left[E_{\varphi|y_o,\mathcal{M},\omega}\left[\varphi\right]\right]\right\} \\
& + V_\omega\left\{E_{\mathcal{M}|y_o,\omega}\left[E_{\varphi|y_o,\mathcal{M},\omega}\left[\varphi\right]\right]\right\} \\
= \quad & \underbrace{E_\omega\left\{E_{\mathcal{M}|y_o,\omega}\left[V_{\varphi|y_o,\mathcal{M},\omega}\left[\varphi\right]\right]\right\}}_{1} + \underbrace{E_\omega\left\{V_{\mathcal{M}|y_o,\omega}\left[E_{\varphi|y_o,\mathcal{M},\omega}\left[\varphi\right]\right]\right\}}_{2} \\
& + \underbrace{V_\omega\left\{E_{\mathcal{M}|y_o,\omega}\left[E_{\varphi|y_o,\mathcal{M},\omega}\left[\varphi\right]\right]\right\}}_{3}.
\end{aligned}
\tag{4.1}
$$

Terms (1) and (2) represent expectations over $p(\omega)$ of within-model variance and between-model variance, respectively. The new weighting variance term (3) reflects the variability of the expected prediction due to uncertainty in the variable ω. All three terms can be evaluated based on the results of the resampling analysis presented above.

Bayesian vs. frequentist interpretation Although mathematically sound, there is a difficulty in interpreting the variance decomposition in Equation 4.1: the expectations and variances over $p(\omega)$ are frequentist properties of the otherwise Bayesian statistics and probabilities, because they do not follow from Bayes' theorem, but from a resampling analysis according to specified statistics. Hence, it might be argued that frequentist confidence intervals should not be mingled with Bayesian credible intervals.

The apparent conflict can be resolved by understanding that the resampling analysis is merely a frequentist tool to make weighting uncertainty visible, i.e. to isolate it from the other two components within-model variance and between-model variance. The total variance could just as well be determined from the traditional Bayesian approach by incorporating the uncertainty in the variable ω into the calculation of Bayesian model evidence to obtain an aggregated model weight.

Note that in the resampling analysis, model weights are treated as uncertain quantities with a probability distribution of their own. Since model weights themselves represent a distribution over the competing models, two levels of statistical distributions arise. These two distributions can be collapsed, and hence one again arrives at a single model weight per model. This supports the hypothesis that both the Bayesian approach, yielding fixed model weights, and the frequentist approach, yielding distributions of weights, can be used to determine the overall predictive uncertainty (Equation 4.1).

The final uncertainty estimates of model-averaged predictions obtained by the two different approaches are expected to roughly coincide if prior information is dominated by the data and a sufficiently large number of resampled data sets is used. In a follow-up study on *Schöniger et al.* (2015b), I will investigate this hypothesis in synthetic and real-world test cases. It should be kept in mind, however, that the mixed confidence and credible intervals obtained from resampling are not necessarily representative if the required assumptions are not fulfilled. Thus, I recommend to perform the Bayesian approach to correctly quantify the overall prediction uncertainty, and to perform the frequentist resampling analysis if the variability in model weights shall be explicitly evaluated as an intermediate result that is otherwise hidden in the Bayesian approach. As demonstrated in the case study of soil-plant modeling, the variability in model weights is of particular relevance for the purpose of model ranking and model selection.

Treatment of input uncertainty The Bayesian approach of aggregated model weights has been implemented for the case of input uncertainty (*Ajami et al.*, 2007; *Rojas et al.*, 2008). Alternatively, input uncertainty could be specifically addressed by the proposed frequentist resampling analysis to obtain an ensemble of model weights. These two options differ in their implicit meaning. If the aggregated Bayesian approach is chosen, the considered source of uncertainty is automatically attributed to the model and will be penalized by BMA in the spirit of parsimony. This approach could be useful, e.g., if input uncertainty varies for the competing models. If input uncertainty is the same for all models, it might be more interesting to find out, how robust the model ranking is against this source of uncertainty. Then, instead of integrating over the uncertain input, the resampling analysis should be performed.

Treatment of measurement noise In the case of measurement noise in the calibration data as a source of uncertainty for model weights, the interpretation of weighting uncertainty becomes trickier. I have claimed that model weights turn uncertain because their outcome varies with different outcomes of measurement error in the data set. The standard BMA routine requires to specify the likelihood function according to the assumptions about the distribution of measurement error. Repeating this analysis for perturbed data means that each time, the likelihood function is moved in order to center it about the respective perturbed value (cf. Figure 2.1c). Since the perturbation is done according to the very same distribution of measurement error, this is equivalent to the convolution of the distribution of measurement error with itself. In the case of a Gaussian likelihood function, this results in a variance twice as large as the measurement

error variance σ_ϵ^2. Thus, to evaluate the Bayesian expression of the total variance (the left side of Equation 4.1), one needs to perform the standard BMA routine based on a likelihood function with a variance of $\sigma^2 = 2\sigma_\epsilon^2$.

Whether, from a methodological viewpoint, the impact of measurement error should be accounted for twice is debatable, since this approach clearly mixes the Bayesian interpretation (measurement values are perceived as fixed values, measurement error is accounted for by the likelihood function) with the frequentist approach (measurement error is accounted for through repeated measurements). It will be tested in a future study whether this mixed approach could still prove useful from a practical viewpoint by analyzing the predictive coverage achieved by mixed confidence/credible intervals of predictions in an independent validation setup.

Summary and implications I have proposed to extend the existing BMA routine to also account for sources of uncertainty for model weights, such as noisy calibration data or uncertain model input. With the concept of resampling, the robustness of model weights and model ranking against the investigated source of uncertainty can be assessed. The application to a case study of soil-plant model selection has shown that model ranking can be highly sensitive to the outcome of random measurement errors in the calibration data set. The proposed extended BMA routine should be used to detect whether this sensitivity is high in any specific application, and to guide further data collection aiming at an increased confidence in model ranking. The proposed resampling analysis further offers the chance to determine a distance in performance between individual models and a theoretical upper limit. Knowing this distance helps to guide further model development.

Beyond the impact on model ranking results, I have claimed that the uncertainty in model weights adds to the total predictive uncertainty. Statistically, the propagation of weighting uncertainty to predictive uncertainty can be developed with the law of total variance. However, difficulties in the interpretation of this variance decomposition arise because frequentist properties are apparently mixed with Bayesian probabilities. I have argued that these difficulties can be resolved by differentiating between two approaches with two distinct goals: the Bayesian approach should be used to correctly quantify the overall predictive uncertainty inherent to the model-averaged predictions. The frequentist resampling approach should be used to make the variability in model weights visible (which would otherwise remain hidden in the Bayesian procedure), and hence to verify that the obtained BMA results are robust and meaningful.

5 Conclusions & Outlook

As explained in the introduction to this thesis, several obstacles to an efficient and meaningful use of BMA exist(ed). Some of these obstacles have been addressed in this thesis by answering the research questions put forward in Chapter 3.

Part I: The necessity to place BMA into a Monte Carlo framework First, I have addressed the question of how to evaluate BMA weights efficiently and accurately by performing a rigorous comparison of existing methods to determine Bayesian model evidence, which is the key ingredient of the BMA equations. I have conducted a first-time benchmarking of mathematical approximations and numerical evaluation schemes against a reference solution under both synthetic and real-world conditions. Results have shown that computationally cheap approximations in form of information criteria yield potentially very inaccurate model weights that do not reflect the true Bayesian trade-off between performance and parsimony. The discrepancy is especially large in the case of nonlinear models. Instead, I recommend to use brute-force Monte Carlo integration whenever feasible in order to obtain an accurate Bayesian model ranking. Solving the BMA equations with Monte Carlo integration is conceptually simple and its implementation is straightforward such that the use of BMA should not be hindered by technical difficulties anymore, given that enough computational power is available. It remains an open question for future investigation whether the so far unacceptable performance of information criteria in approximating Bayesian model evidence could be improved in order to create a computationally less demanding but accurate alternative to Monte Carlo integration. A specific focus should be laid on how to adequately encode model complexity, because this aspect is currently causing the information criteria to fail.

Part II: The benefit of performing a model justifiability analysis Second, I have investigated the mechanism of BMA that balances performance with parsimony. The proposed model justifiability analysis allows a better understanding of this trade-off and facilitates the interpretation of BMA results. It answers the question, whether the complexity in the considered models could possibly be justified with the data available for calibration. If the complexity of a model cannot be justified, BMA might favor simpler models not because they seem more realistic, but because they are better justifiable

in light of limited data. The proposed model justifiability analysis can be applied to arbitrary models and data and comes at little additional costs if the standard BMA analysis is already performed in a Monte Carlo framework, which I highly recommend based on the findings from the first part of this thesis. The proposed two-step procedure has yielded valuable insights into the required or justified complexity of groundwater models for a synthetic sandbox aquifer. Transferring this methodology to field-scale applications poses a challenge to be addressed in future studies.

Part III: The importance of assessing the robustness of model weights Third, I have addressed the question of how robust and representative model ranking results are. Therefore, I have extended the standard BMA routine to also account for sources of uncertainty for model weights, such as measurement noise in calibration data or in input data, or conceptual uncertainty in boundary conditions. Since assumptions on the distribution of measurement errors in the calibration data set are already part of the standard BMA analysis, it is advisable to assess the variability in model weights due to these very assumptions in any BMA application. Results from a case study of soil-plant model selection have confirmed the hypothesis that model weights can be highly sensitive to the outcome of random measurement errors, which compromises the confidence in model ranking. As a highly beneficial side product, the proposed analysis provides an indication how well individual models perform as compared to a theoretical optimum. Testing the proposed extended BMA routine on uncertain forcings or boundary conditions remains open for future studies. The extended BMA framework is further a solid basis for the optimal design of future data collection campaigns, because it can point the modeler towards those measurements of input or output data that yield a most decisive and most confident model ranking. The definition of utility functions that also account for the negative impact of weighting uncertainty will be in the focus of future research. It should further be investigated in synthetic and real-world case studies, how relevant the contribution of weighting uncertainty to the overall predictive uncertainty is.

Remaining issues for future research These three distinct contributions have paved the ground for using BMA as a general-purpose multi-modeling framework that can consistently handle sources of uncertainty in models, parameters, inputs, and data. The proposed extended BMA routine is a valuable addition to existing integrated modeling approaches, because it is able to explicitly evaluate the variability in model weights due to any one of these sources of uncertainty. One link is, however, still missing to

fully establish BMA as an integrated modeling framework especially for hydrosystem modeling. Equipping each process-based model with its own statistical error model to reduce the bias in predictions has so far not been explicitly attempted within BMA. This seems to be an adjustment long overdue, since acknowledging the existence of model structural errors is the main motivation to use multi-model frameworks in the first place. It is expected that incorporating model structural errors into the BMA routine will further improve the predictive coverage of BMA-weighted predictions and will allow for a potentially more realistic estimation of model parameters and their attached uncertainty.

While this thesis has advanced Bayesian multi-modeling with regard to its technical implementation, its interpretation, and its reliability, some of the listed obstacles persist. Especially the population of the model space remains an unresolved challenge. Some of the analysis tools presented in this thesis can be used to assess the degree of (dis)similarity and hence the identifiability of alternative models: the model confusion matrix assesses this distance in an a priori sense (without considering the actually observed data), and the distance from the theoretically optimal model is a measure in relation to the observed data. Knowledge about the distance between competing models and the absolute distance from a theoretical optimum can give some insight about the structure of the currently sampled model space.

The subjective choice of prior weights for the models in the set will also remain in the focus of BMA criticism. But as long as there is a lack of better suited alternatives, I believe this obstacle should not prevent us from using BMA. In the end, BMA is an objective, coherent and transparent approach that starts from a subjective (but explicitly stated) viewpoint of the modeler, and this is certainly much better scientific practice than starting from a subjective position (not necessarily made explicit) and moving forward based on further subjective decisions. Equipped with the tools developed within this thesis and anticipating further theoretical development as suggested here, BMA will help to advance hydrosystem modeling in various aspects, such as system understanding, uncertainty quantification and forecasting skill.

Bibliography

Abramowitz, G., and H. Gupta (2008), Toward a model space and model independence metric, *Geophysical Research Letters*, *35*(5), doi:10.1029/2007gl032834.

Ajami, N. K., and C. Gu (2010), Complexity in microbial metabolic processes in soil nitrogen modeling: A case for model averaging, *Stochastic Environmental Research and Risk Assessment*, *24*(6), 831–844, doi:10.1007/s00477-010-0381-4.

Ajami, N. K., Q. Y. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resources Research*, *43*(1), doi:10.1029/2005wr004745.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki, pp. 367–281.

Asmis, E. (1984), *Epicurus' Scientific Method*, Cornell studies in classical philology, Cornell University Press, Ithaca, NY.

Berg, S. J., and W. A. Illman (2011), Capturing aquifer heterogeneity: Comparison of approaches through controlled sandbox experiments, *Water Resources Research*, *47*(9), doi:10.1029/2011wr010429.

Beven, K., and A. Binley (1992), The future of distributed models - model calibration and uncertainty prediction, *Hydrological Processes*, *6*(3), 279–298, doi:10.1002/hyp.3360060305.

Box, G., and G. Tiao (1973), *Bayesian inference in statistical analysis*, Addison-Wesley, Reading, Massachusetts.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997), Model selection: An integral part of inference, *Biometrics*, *53*(2), 603–618, doi:10.2307/2533961.

Burnham, K. P., and D. R. Anderson (2003), *Model selection and multimodel inference, a practical information theoretic approach*, chap. 1, 2nd [corr. print.] ed., Springer, New York.

Burnham, K. P., and D. R. Anderson (2004), Multimodel inference - understanding AIC and BIC in model selection, *Sociological Methods and Research*, *33*(2), 261–304, doi:10.1177/0049124104268644.

Chamberlin, T. C. (1890), The method of multiple working hypotheses, *Science*, *15*(366), 92–96.

Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resources Research*, *47*(9), doi:10.1029/2010wr009827.

Davison, A. C., and D. V. Hinkley (1997), *Bootstrap methods and their application*, chap. 4.2, 1st ed., Cambridge University Press.

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

Drton, M., B. Sturmfels, and S. Sullivant (2009), *Bayesian Integrals*, *Oberwolfach Seminars*, vol. 39, chap. 5, pp. 105–121, Birkhäuser Basel.

Elshall, A. S., and F. T. C. Tsai (2014), Constructive epistemic modeling of groundwater flow with geological structure and boundary condition uncertainty under the Bayesian paradigm, *Journal of Hydrology*, *517*, 105–119, doi:10.1016/j.jhydrol.2014.05.027.

Elsheikh, A. H., M. F. Wheeler, and I. Hoteit (2013), Nested sampling algorithm for subsurface flow model selection, uncertainty quantification, and nonlinear calibration, *Water Resources Research*, *49*(12), 8383–8399, doi:10.1002/2012wr013406.

Fienen, M. N., T. Clemo, and P. K. Kitanidis (2008), An interactive Bayesian geostatistical inverse protocol for hydraulic tomography, *Water Resources Research*, *44*(12), doi:10.1029/2007wr006730.

Fisher, R. A. (1922), On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society Series A*, *222*, 309.

Foglia, L., S. W. Mehl, M. C. Hill, and P. Burlando (2013), Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland, *Water Resources Research*, *49*(1), 260–282, doi:10.1029/2011wr011779.

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

George, E. I. (2010), *Dilution priors: Compensating for model space redundancy*, pp. 158–165, Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown, Institute of Mathematical Statistics.

Goodarzi, E., M. Ziaei, and L. T. Shui (2013), *Introduction to risk and uncertainty in hydrosystem engineering*, *Topics in Safety, Risk, Reliability and Quality*, vol. 22, Springer Science & Business Media.

Gull, S. F. (1988), Bayesian inductive inference and maximum entropy, *Maximum Entropy and Bayesian Methods in Science and Engineering*, *1*, 53–74.

Hall, J., and D. Solomatine (2008), A framework for uncertainty analysis in flood risk management decisions, *International Journal of River Basin Management*, *6*(2), 85–98, doi:10.1080/15715124.2008.9635339.

Hammersley, J. M., and D. C. Handscomb (2013), *Monte Carlo methods*, Monographs on Applied Probability and Statistics, Chapman and Hall Ltd.

Hill, M. C., and C. R. Tiedeman (2007), *Effective groundwater model calibration, with analysis of data, sensitivities, predictions, and uncertainty*, Wiley-Interscience, Hoboken, NJ.

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

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro (2004), Bayesian phylogenetic model selection using Reversible Jump Markov Chain Monte Carlo, *Molecular Biology and Evolution*, *21*(6), 1123–1133, doi:10.1093/molbev/msh123.

Hutter, M. (2006), On generalized computable universal priors and their convergence, *Theoretical Computer Science*, *364*(1), 27–41, doi:http://dx.doi.org/10.1016/j.tcs.2006.07.039.

Illman, W. A., J. Zhu, A. J. Craig, and D. Yin (2010), Comparison of aquifer characterization approaches through steady state groundwater model validation: A controlled laboratory sandbox study, *Water Resources Research*, *46*(4), doi:10.1029/2009wr007745.

Illman, W. A., S. J. Berg, and Z. Zhao (2015), Should hydraulic tomography data be interpreted using geostatistical inverse modeling? A laboratory sandbox investigation, *Water Resources Research*, *51*(5), 3219–3237, doi:10.1002/2014WR016552.

Jaynes, E. T. (1985), Bayesian methods: General background, in *In Maximum Entropy and Bayesian Methods in Applied Statistics*, edited by J. H. Justice, pp. 1 – 25, Cambridge University Press, Cambridge.

Jefferys, W. H., and J. O. Berger (1992), Ockham's razor and Bayesian analysis, *American Scientist*, pp. 64–72.

Jeffreys, H. (1939), *Theory of probability*, chap. 5, 1st ed., Clarendon Press, Oxford.

Jeffreys, H. (1961), *Theory of probability*, 3rd ed., Oxford University Press.

Kadane, J. B., and N. A. Lazar (2004), Methods and criteria for model selection, *Journal of the American Statistical Association*, *99*(465), 279–290, doi:10.1198/016214504000000269.

Kass, R. E., and A. E. Raftery (1995), Bayes factors, *Journal of the American Statistical Association*, *90*(430), 773–795, doi:10.2307/2291091.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resources Research*, *42*(3), doi:10.1029/2005WR004368.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resources Research*, *42*(3), doi:10.1029/2005WR004376.

Kitanidis, P. K. (1986), Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resources Research*, *22*(4), 499–507, doi:10.1029/WR022i004p00499.

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *Journal of Hydrology*, *331*(1-2), 161–177, doi:10.1016/j.jhydrol.2006.05.010.

Liu, X. Y., M. A. Cardiff, and P. K. Kitanidis (2010), Parameter estimation in nonlinear environmental problems, *Stochastic Environmental Research and Risk Assessment*, *24*(7), 1003–1022, doi:10.1007/s00477-010-0395-y.

Liu, Y. Q., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resources Research*, *43*(7), doi:10.1029/2006wr005756.

Lötgering-Lin, O., A. Schöniger, W. Nowak, and J. Gross (2015), Choosing between thermodynamic models with Bayesian model selection: A case study on the calculation of viscosities via entropy scaling, *manuscript in preparation*.

Lu, D., M. Ye, and S. P. Neuman (2011), Dependence of Bayesian model selection criteria and Fisher information matrix on sample size, *Mathematical Geosciences*, *43*(8), 971–993, doi:10.1007/s11004-011-9359-0.

Lu, D., M. Ye, S. P. Neuman, and L. Xue (2012), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, *Advances in Water Resources*, *35*, 69–82, doi:10.1016/j.advwatres.2011.10.007.

Lu, D., M. Ye, P. D. Meyer, G. P. Curtis, X. Q. Shi, X. F. Niu, and S. B. Yabusaki (2013), Effects of error covariance structure on estimation of model averaging weights and predictive performance, *Water Resources Research*, *49*(9), 6029–6047, doi:10.1002/Wrcr.20441.

MacKay, D. J. C. (1992), Bayesian interpolation, *Neural Computation*, *4*, 415–447, doi:10.1162/neco.1992.4.3.415.

Morales-Casique, E., S. P. Neuman, and V. V. Vesselinov (2010), Maximum likelihood Bayesian averaging of airflow models in unsaturated fractured tuff using Occam and variance windows, *Stochastic Environmental Research and Risk Assessment*, *24*(6), 863–880, doi:10.1007/s00477-010-0383-2.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, and D. A. Stainforth (2004), Quantification of modelling uncertainties in a large ensemble of climate change simulations, *Nature*, *430*(7001), 768–772, doi:10.1038/Nature02771.

Najafi, M. R., H. Moradkhani, and I. W. Jung (2011), Assessing the uncertainties of hydrologic model selection in climate change impact studies, *Hydrological Processes*, *25*(18), 2814–2826, doi:10.1002/Hyp.8043.

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

Neuman, S. P., L. Xue, M. Ye, and D. Lu (2012), Bayesian analysis of data-worth considering model and parameter uncertainties, *Advances in Water Resources*, *36*, 75–85, doi:10.1016/j.advwatres.2011.02.007.

Nowak, W. (2010), Measures of parameter uncertainty in geostatistical estimation and geostatistical optimal design, *Mathematical Geosciences*, *42*(2), 199–221, doi:10.1007/s11004-009-9245-1.

Palmer, T. N., and J. Raisanen (2002), Quantifying the risk of extreme seasonal precipitation events in a changing climate, *Nature*, *415*(6871), 512–514.

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

Raftery, A. E. (1995), Bayesian model selection in social research, *Sociological Methodology 1995*, *25*, 111–163, doi:10.2307/271063.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Monthly Weather Review*, *133*(5), 1155–1174, doi:10.1175/mwr2906.1.

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

Refsgaard, J. C., S. Christensen, T. O. Sonnenborg, D. Seifert, A. L. Hojberg, and L. Troldborg (2012), Review of strategies for handling geological uncertainty in groundwater flow and transport modeling, *Advances in Water Resources*, *36*, 36–50, doi:10.1016/j.advwatres.2011.04.006.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resources Research*, *46*(5), doi:10.1029/2009wr008328.

Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resources Research*, *44*(12), doi:10.1029/2008wr006908.

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

Rojas, R., L. Feyen, O. Batelaan, and A. Dassargues (2010), On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling, *Water Resources Research*, *46*(8), doi:10.1029/2009wr008822.

Samaniego, L., R. Kumar, and S. Attinger (2010), Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, *Water Resources Research*, *46*(5), doi:10.1029/2008wr007327.

Schöniger, A., T. Wöhling, L. Samaniego, and W. Nowak (2014), Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence, *Water Resources Research*, *50*(12), 9484–9513, doi:10.1002/2014WR016062.

Schöniger, A., W. A. Illman, T. Wöhling, and W. Nowak (2015a), Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection, *Journal of Hydrology*, *531*(1), 96–110, doi:10.1016/j.jhydrol.2015.07.047.

Schöniger, A., T. Wöhling, and W. Nowak (2015b), A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking, *Water Resources Research*, *51*(9), 7524–7546, doi:10.1002/2015WR016918.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, *6*(2), 461–464, doi:10.1214/aos/1176344136.

Seifert, D., T. O. Sonnenborg, J. C. Refsgaard, A. L. Hojberg, and L. Troldborg (2012), Assessment of hydrological model predictive ability given multiple conceptual geological models, *Water Resources Research*, *48*(6), doi:10.1029/2011wr011149.

Singh, A., S. Mishra, and G. Ruskauff (2010), Model averaging techniques for quantifying conceptual model uncertainty, *Ground Water*, *48*(5), 701–715, doi:10.1111/j.1745-6584.2009.00642.x.

Skilling, J. (2006), Nested sampling for general Bayesian computation, *Bayesian Analysis*, *1*(4), 833–859.

Troldborg, M., W. Nowak, N. Tuxen, P. L. Bjerg, R. Helmig, and P. J. Binning (2010), Uncertainty evaluation of mass discharge estimates from a contaminated site using a fully Bayesian framework, *Water Resources Research*, *46*(12), doi:10.1029/2010wr009227.

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, doi:10.1007/s00477-010-0382-3.

Tsai, F. T. C., and A. S. Elshall (2013), Hierarchical Bayesian model averaging for hydrostratigraphic modeling: Uncertainty segregation and comparative evaluation, *Water Resources Research*, *49*(9), 5520–5536, doi:10.1002/Wrcr.20428.

Tsai, F. T.-C., and X. B. Li (2008), Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window, *Water Resources Research*, *44*(9), doi:10.1029/2007wr006576.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resources Research*, *44*(12), doi:10.1029/2007wr006720.

Wöhling, T., and J. A. Vrugt (2008), Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models, *Water Resources Research*, *44*(12), doi:10.1029/2008wr007154.

Wöhling, T., A. Schöniger, S. Gayler, and W. Nowak (2015), Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction, *Water Resources Research*, *51*(4), 2825–2846, doi:10.1002/2014wr016292.

Xue, L., D. Zhang, A. Guadagnini, and S. P. Neuman (2014), Multimodel Bayesian analysis of groundwater data worth, *Water Resources Research*, *50*(11), 8481–8496, doi:10.1002/2014wr015503.

Ye, M., S. P. Neuman, P. D. Meyer, and K. Pohlmann (2005), Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff, *Water Resources Research*, *41*(12), doi:10.1029/2005wr004260.

Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resources Research*, *44*(3), doi:10.1029/2008wr006803.

Ye, M., K. F. Pohlmann, J. B. Chapman, G. M. Pohll, and D. M. Reeves (2010), A model-averaging method for assessing groundwater conceptual model uncertainty, *Ground Water*, *48*(5), 716–28, doi:10.1111/j.1745-6584.2009.00633.x.

List of Publications

1. Schöniger, A., Wöhling, T., Samaniego, L., & Nowak, W. (2014). Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. Water Resources Research, 50(12), 9484-9513.
   DOI: 10.1002/2014WR016062

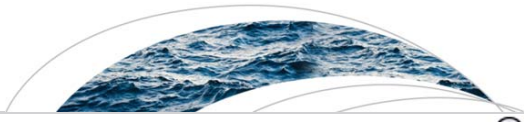
2. Schöniger, A., Illman, W. A., Wöhling, T., & Nowak, W. (2015). Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. Journal of Hydrology, 531(1), 96-110.
   DOI: 10.1016/j.jhydrol.2015.07.047

3. Schöniger, A., Wöhling, T., & Nowak, W. (2015). A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. Water Resources Research, 51(9), 7524-7546.
   DOI: 10.1002/2015WR016918

A Publications

AGU PUBLICATIONS

Water Resources Research



RESEARCH ARTICLE
10.1002/2014WR016062

Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence

Anneli Schöniger[1], Thomas Wöhling[2,3], Luis Samaniego[4], and Wolfgang Nowak[5]

[1]Center for Applied Geoscience, University of Tübingen, Tübingen, Germany, [2]Water and Earth System Science (WESS) Competence Cluster, University of Tübingen, Tübingen, Germany, [3]Lincoln Environmental Research, Lincoln Agritech, Hamilton, New Zealand, [4]Department Computational Hydrosystems, Helmholtz-Zentrum for Environmental Research—UFZ, Leipzig, Germany, [5]Institute for Modelling Hydraulic and Environmental Systems (LS3)/SimTech, University of Stuttgart, Stuttgart, Germany

Key Points:
- The choice of BME evaluation method influences the outcome of model ranking
- Out of the ICs, the KIC@MAP is the most consistent one
- For reliable model selection, there is still no alternative to numerical methods

Correspondence to:
A. Schöniger,
anneli.schoeniger@uni-tuebingen.de

Citation:
Schöniger, A., T. Wöhling, L. Samaniego, and W. Nowak (2014), Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence, *Water Resour. Res., 50*, 9484–9513, doi:10.1002/2014WR016062.

Received 27 JUN 2014
Accepted 30 OCT 2014
Accepted article online 4 NOV 2014
Published online 19 DEC 2014

Abstract Bayesian model selection or averaging objectively ranks a number of plausible, competing conceptual models based on Bayes' theorem. It implicitly performs an optimal trade-off between performance in fitting available data and minimum model complexity. The procedure requires determining Bayesian model evidence (BME), which is the likelihood of the observed data integrated over each model's parameter space. The computation of this integral is highly challenging because it is as high-dimensional as the number of model parameters. Three classes of techniques to compute BME are available, each with its own challenges and limitations: (1) Exact and fast analytical solutions are limited by strong assumptions. (2) Numerical evaluation quickly becomes unfeasible for expensive models. (3) Approximations known as information criteria (ICs) such as the AIC, BIC, or KIC (Akaike, Bayesian, or Kashyap information criterion, respectively) yield contradicting results with regard to model ranking. Our study features a theory-based intercomparison of these techniques. We further assess their accuracy in a simplistic synthetic example where for some scenarios an exact analytical solution exists. In more challenging scenarios, we use a brute-force Monte Carlo integration method as reference. We continue this analysis with a real-world application of hydrological model selection. This is a first-time benchmarking of the various methods for BME evaluation against true solutions. Results show that BME values from ICs are often heavily biased and that the choice of approximation method substantially influences the accuracy of model ranking. For reliable model selection, bias-free numerical methods should be preferred over ICs whenever computationally feasible.

1. Introduction

The idea of model validation is to objectively scrutinize a model's ability to reproduce an observed data set and then to falsify the hypothesis that this model is a good representation for the system under study [*Popper*, 1959]. If this hypothesis cannot be rejected, the model may be considered for predictive purposes. Modelers have been encouraged for centuries to create multiple such working hypotheses instead of limiting themselves to the subjective choice of a single conceptual representation, therewith avoiding the "dangers of parental affection for a favorite theory" [*Chamberlin*, 1890]. These dangers include a significant underestimation of predictive uncertainty due to the neglected conceptual uncertainty (uncertainty in the choice of a most adequate representation of a system). Recognizing conceptual uncertainty as a main contribution to overall predictive uncertainty [e.g., *Burnham and Anderson*, 2003; *Gupta et al.*, 2012; *Clark et al.*, 2011; *Refsgaard et al.*, 2006] makes model selection an "integral part of inference" [*Buckland et al.*, 1997]. The quantification of conceptual uncertainty is of importance in a variety of scientific disciplines, e.g., in climate change modeling [*Murphy et al.*, 2004; *Najafi et al.*, 2011], weather forecasting [*Raftery et al.*, 2005], hydrogeology [*Rojas et al.*, 2008; *Poeter and Anderson*, 2005; *Ye et al.*, 2010a], geostatistics [*Neuman*, 2003; *Ye et al.*, 2004], vadose zone hydrology [*Wöhling and Vrugt*, 2008], and surface hydrology [*Ajami et al.*, 2007; *Vrugt and Robinson*, 2007; *Renard et al.*, 2010], to name only a few selected examples from the field of water resources.

Different strategies have been proposed to develop alternative conceptual models, assess their strengths and weaknesses, and to test their predictive ability. Bayesian model averaging (BMA) [*Hoeting et al.*, 1999] is a formal statistical approach which allows comparing alternative conceptual models, testing their adequacy, combining their predictions into a more robust output estimate, and quantifying the contribution of

conceptual uncertainty to the overall prediction uncertainty. The BMA approach is based on Bayes' theorem, which combines a prior belief about the adequacy of each model with its performance in reproducing a common data set. It yields model weights that represent posterior probabilities for each model to be the best one from the set of proposed alternative models. Based on the weights, it allows for a ranking and quantitative comparison of the competing models. Hence, BMA can be understood as a Bayesian hypothesis testing framework, merging the idea of classical hypothesis testing with the ability to test several alternative models against each other in a probabilistic way. The principle of parsimony or "Occam's razor" [e.g., *Angluin and Smith*, 1983] is implicitly followed by Bayes' theorem, such that the posterior model weights reflect a compromise between model complexity and goodness of fit (also known as the bias-variance trade-off [*Geman et al*., 1992]). BMA has been adopted in many different fields of research, e.g., sociology [*Raftery*, 1995], ecology [*Link and Barker*, 2006], hydrogeology [*Li and Tsai*, 2009], or contaminant hydrology [*Troldborg et al*., 2010], indicating the general need for such a systematic model selection procedure.

The drawback of BMA is, however, that it involves the evaluation of a quantity called Bayesian model evidence (BME). This integral over a model's parameter space typically cannot be computed analytically, while numerical solutions come at the price of high computational costs. Various authors have suggested and applied different approximations to the analytical BMA equations to render the procedure feasible. *Neuman* [2003] proposes a Maximum Likelihood Bayesian Model Averaging approach (MLBMA), which reduces computational effort by evaluating Kashyap's information criterion (KIC) for the most likely parameter set instead of integrating over the whole parameter space. This is especially compelling for high-dimensional applications (i.e., models with many parameters). If prior knowledge about the parameters is not available or vague, a further simplification leads to the Bayesian information criterion or Schwarz' information criterion (BIC) [*Schwarz*, 1978; *Raftery*, 1995]. The Akaike information criterion (AIC) [*Akaike*, 1973] originates from information theory and is frequently applied in the context of BMA in social research [*Burnham and Anderson*, 2003] for its ease of implementation. Previous studies have revealed that these information criteria (IC) differ in the resulting posterior model weights or even in the ranking of the models [*Poeter and Anderson*, 2005; *Ye et al*., 2008, 2010a, 2010b; *Tsai and Li*, 2010, 2010; *Singh et al*., 2010; *Morales-Casique et al*., 2010; *Foglia et al*., 2013]. This implies that they do not reflect the true Bayesian trade-off between performance and complexity, but might produce an arbitrary trade-off which is not supported by Bayesian theory and cannot provide a reliable basis for Bayesian model selection. *Burnham and Anderson* [2004] conclude that ". . . many reported studies are not appropriate as a basis for inference about which criterion should be used for model selection with real data." The work of *Lu et al*. [2011] has been a first step into clarifying the so far contradictory results by comparing the KIC and the BIC against a Markov chain Monte Carlo (MCMC) reference solution for a synthetic geostatistical application.

Our study aims to advance this endeavor by rigorously assessing and comparing a more comprehensive set of nine different methods to evaluate BME. In specific, we will highlight their theoretical derivation, computational effort, and approximation accuracy. As representatives of mathematical approximations, we consider the AIC, AICc (bias-corrected AIC), and BIC in our comparison. We further include the KIC evaluated at the maximum likelihood parameter estimate (KIC@MLE) as introduced in MLBMA, and an alternative formulation that is evaluated at the maximum a posteriori parameter estimate instead (KIC@MAP). We also consider three types of Monte Carlo integration techniques (simple Monte Carlo integration, MC; MC integration with importance sampling, MC IS; MC integration with posterior sampling, MC PS) and a very recent approach called nested sampling (NS) as representatives of numerical methods. By pointing out and comparing the important features and assumptions of these mostly well-known techniques, we are able to argue which methods are truly suitable for BME evaluation, and which ones are suspected to yield inaccurate results. We then present a simplistic synthetic, linear test case where an exact analytical expression for BME exists. With this first-time benchmarking of the different BME evaluation methods against the true solution, we close a significant gap in the model selection literature.

The controlled setup in the simplistic example allows us to systematically investigate the factors which influence the value of BME and the approximation thereof by the nine featured evaluation methods. The two main factors investigated are (1) the size of the data set which determines the "seriousness" of the goodness of fit rating, and (2) the shape of the parameter prior which characterizes the robustness of a model. In a second step, we assess the performance of the different methods when confronted with low-dimensional nonlinear models. In this more challenging scenario of the synthetic example, no analytical solution to

compute BME exists. We therefore generate a reference solution by brute-force MC integration, after having proven its suitability as reference solution in the linear case. In a third step, we present a real-world application of hydrological model selection. We chose this application such that the model selection task is still relatively simple and unambiguous. Even in this case the deficiencies of some of the evaluation methods become apparent. Our systematic investigation of methods to determine BME takes an important next step toward robust model selection in agreement with Bayes' theorem, *heaving it up on solid ground*.

We summarize the statistical framework of BMA in section 2 and discuss assets and drawbacks of the available techniques to determine BME in section 3. In section 4, we present the first-time benchmarking of the featured methods on the simplistic test case. Section 5 compares the approximation performance in a real-world hydrological model selection problem. We summarize our findings and formulate recommendations on which methods to use for reliable model selection in even more complex situations in section 6.

2. Bayesian Model Averaging Framework

We formulate the BMA equations according to *Hoeting et al*. [1999]. All probabilities and statistics are implicitly conditional on the set of considered models. While the suite of models is a subjective choice that lies in the responsibility of the modeler, it is the starting point for a systematic procedure to account for model uncertainty based on objective likelihood measures.

Let us consider N_m plausible, competing models M_k. The posterior predictive distribution of a quantity of interest φ given the vector of observed data $\mathbf{y}_o$ can be expressed as:

$$p(\varphi|\mathbf{y}_o)=\sum_{k=1}^{N_m} p(\varphi|\mathbf{y}_o, M_k)P(M_k|\mathbf{y}_o) \tag{1}$$

with $p(\cdot|\mathbf{y}_o)$ representing a conditional probability distribution and $P(M_k|\mathbf{y}_o)$ being discrete posterior model weights. The weights can be interpreted as the Bayesian probability of the individual models to be the best representation of the system from the set of considered models.

The model weights are given by Bayes' theorem:

$$P(M_k|\mathbf{y}_o)=\frac{p(\mathbf{y}_o|M_k)P(M_k)}{\sum_{i=1}^{N_m} p(\mathbf{y}_o|M_i)P(M_i)}, \tag{2}$$

with the prior probability (or rather subjective model credibility) $P(M_k)$ that model M_k could be the best one (the most plausible, adequate, and consistent one) in the set of models *before* any observed data have been considered. A "reasonable, neutral choice" [*Hoeting et al*., 1999] could be equally likely priors $P(M_k)=1/N_m$ if there is little prior knowledge about the assets of the different models under consideration. The denominator in equation (2) is the normalizing constant of the posterior distribution of the models. It is easily obtained by determination of the individual weights. It could even be neglected, since all model weights are normalized by the same constant, so that the ranking of the individual models against each other is fully defined by the proportionality:

$$P(M_k|\mathbf{y}_o) \propto p(\mathbf{y}_o|M_k)P(M_k). \tag{3}$$

$p(\mathbf{y}_o|M_k)$ represents the BME term as introduced in section 1 and is also referred to as *marginal likelihood* or *prior predictive* because it quantifies the likelihood of the observed data based on the prior distribution of the parameters:

$$p(\mathbf{y}_o|M_k)=\int_{\mathcal{U}_k} p(\mathbf{y}_o|M_k, \mathbf{u}_k)p(\mathbf{u}_k|M_k)d\mathbf{u}_k, \tag{4}$$

where $\mathbf{u}_k$ denotes the vector of parameters of model M_k with dimension equal to the number $N_{p,k}$ of parameters, $\mathcal{U}_k$ is the corresponding parameter space, and $p(\mathbf{u}_k|M_k)$ denotes their prior distribution. $p(\mathbf{y}_o|M_k, \mathbf{u}_k)$ is the likelihood or probability of the parameter set $\mathbf{u}_k$ of model M_k to have generated the observed data set. The BME term can either be evaluated via integration over the full parameter space $\mathcal{U}_k$ (equation (4), referred to as Bayesian integral by *Kass and Raftery* [1995]), or via the posterior probability distribution of the parameters $p(\mathbf{u}_k|M_k, \mathbf{y}_o)$ by rewriting Bayes' theorem with respect to the parameter distribution (instead of the model distribution, equation (2)):

$$p(\mathbf{u}_k|M_k,\mathbf{y}_o)=\frac{p(\mathbf{y}_o|M_k,\mathbf{u}_k)p(\mathbf{u}_k|M_k)}{p(\mathbf{y}_o|M_k)}. \quad (5)$$

$p(\mathbf{y}_o|M_k)$ acts as a model-specific normalizing constant for the posterior of the parameters $p(\mathbf{u}_k|M_k,\mathbf{y}_o)$. As a matter of fact, evaluating $p(\mathbf{y}_o|M_k)$ for any given model is a major nuisance in Bayesian updating, and MCMC methods have been developed with the goal to entirely avoid its evaluation. However, in order to evaluate BME, this normalizing constant has to be determined, which is the challenge addressed in the current study. Rearranging equation (5) yields the alternative formulation for equation (4):

$$p(\mathbf{y}_o|M_k)=\frac{p(\mathbf{y}_o|M_k,\mathbf{u}_k)p(\mathbf{u}_k|M_k)}{p(\mathbf{u}_k|M_k,\mathbf{y}_o)}. \quad (6)$$

MacKay [1992] refers to the twofold evaluation of Bayes' theorem (equations (2) and (4) or (6)) as the "two levels of inference" in Bayesian model averaging: the first level is concerned with finding the posterior distribution of the models, the second level with finding the posterior distribution of each model's parameters (or rather its normalizing constant).

The integration over the full parameter space in equation (4) can be an exhaustive calculation, especially for high-dimensional parameter spaces $\mathcal{U}_k$. The alternative of computing the posterior distribution of the parameters (defining the "calibrated" parameter space, equation (5)) is similarly demanding in high-dimensional applications. Analytical solutions are available only under strongly limiting assumptions. In general, mathematical approximations or numerical methods have to be drawn upon instead. We discuss and compare the nine different methods to compute BME in the following section and assess their accuracy in section 4.

3. Available Techniques to Determine BME

We will adopt the notation of *Kass and Raftery* [1995] for equation (4):

$$I_k=p(\mathbf{y}_o|M_k)=\int_{\mathcal{U}_k} p(\mathbf{y}_o|M_k,\mathbf{u}_k)p(\mathbf{u}_k|M_k)d\mathbf{u}_k \quad (7)$$

and denote any approximation to the true BME value I_k as $\hat{I}_k$. After explaining two formulations of the analytical solution in detail in section 3.1, we examine mathematical approximations in the form of ICs in section 3.2. Finally, we discuss assets and drawbacks of selected numerical evaluation methods in section 3.3 and summarize our preliminary findings from this theoretical comparison in section 3.4. All BME evaluation methods featured in this study are listed with their underlying assumptions in Table 4. All approximation methods (i.e., the nine nonanalytical approaches) follow equation (7) to evaluate BME. We do not use equation (6) here, since typically for medium to highly parameterized applications, the multivariate probability density of posterior parameter realizations cannot be estimated. Knowing the posterior parameter distribution up to its normalizing constant (as in MCMC methods, see section 3.3.3) does not suffice here since the normalizing constant is actually the targeted quantity itself.

3.1. Analytical Solution

The Bayesian integral or BME I_k for model M_k can be evaluated analytically for exponential family distributions with conjugate priors [see e.g., *DeGroot*, 1970]. Thus, analytical solutions for BME are available, if the observed data $\mathbf{y}_o$ are measurements of the model parameters $\mathbf{u}_k$ or a linear function thereof and a conjugate prior (i.e., the prior parameter distribution is in the same family as the posterior parameter distribution) exists. This is generally not the case in realistic applications. However, we will briefly outline the analytical solution to BME under these restrictive and simple conditions, before we discuss other evaluation methods that are not limited by these strong assumptions in sections 3.2 and 3.3.

We will focus here on the special case of a linear model M_k with a linear model operator $\mathbf{H}_k$ relating multi-Gaussian parameters $\mathbf{u}_k$ to multivariate Gaussian distributed variables $\mathbf{y}_k$:

$$M_k: \mathbf{y}_k=\mathbf{H}_k\mathbf{u}_k. \quad (8)$$

The prior parameter distribution is defined as a normal distribution $p(\mathbf{u}_k)\sim\mathcal{N}(\overline{\mathbf{u}}_k,\mathbf{C}_{uu})$ with the prior mean $\overline{\mathbf{u}}_k$ and the covariance matrix $\mathbf{C}_{uu}$. For simplicity of notation, the index k is dropped from the notation for the parameter covariance matrix.

The residuals $\epsilon = \mathbf{y}_o - \mathbf{y}_k$ signify any type of error associated with the data set and the models, e.g., measurement errors and model errors. Here we assume the models to be perfect (free of model errors) and only measurement errors to be relevant, and adopt a Gaussian model $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ with a diagonal matrix $\mathbf{R}$ representing the covariance matrix for uncorrelated measurement errors. This results in a Gaussian likelihood function $p(\mathbf{y}_o|M_k, \mathbf{u}_k) \sim \mathcal{N}(\mathbf{y}_k, \mathbf{R})$. Using the theory of linear uncertainty propagation [e.g., *Schweppe*, 1973] and the stated assumptions, BME can be directly evaluated for any given data set $\mathbf{y}_o$ from the Gaussian distribution:

$$I_k = p(\mathbf{y}_o|M_k) \sim \mathcal{N}\left(\mathbf{H}_k \bar{\mathbf{u}}_k, \mathbf{C}_{yy} + \mathbf{R}\right), \tag{9}$$

with $\mathbf{C}_{yy} = \mathbf{H}_k \mathbf{C}_{uu} \mathbf{H}_k^T$.

As an alternative way to determine BME analytically, the posterior distribution of the parameters can be derived since the Gaussian distribution family is self-conjugate [*Box and Tiao*, 1973]. In general, the likelihood $\mathcal{L}$ of the observed data given the prior parameter space of model M_k can be written as a function of the parameters $\mathcal{L}(M_k, \mathbf{u}_k|\mathbf{y}_o)$ [*Fisher*, 1922]. Note that the likelihood function is not necessarily a proper probability density function with respect to $\mathbf{u}_k$, because it does not necessarily integrate to one. With the assumptions of Gaussian measurement noise and a linear model, the likelihood can be expressed as a Gaussian function of the parameters $\mathcal{L}(M_k, \mathbf{u}_k|\mathbf{y}_o) \sim \mathcal{N}(\hat{\mathbf{u}}_k, \mathbf{C}_{\hat{u}\hat{u}})$ with $\hat{\mathbf{u}}_k = \left(\mathbf{H}_k^T \mathbf{H}_k\right)^{-1} \mathbf{H}_k^T \mathbf{y}_o$ and $\mathbf{C}_{\hat{u}\hat{u}} = \left[\mathbf{H}_k^T \mathbf{R}^{-1} \mathbf{H}_k\right]^{-1}$. The mean of the distribution, $\hat{\mathbf{u}}_k$, is the maximum likelihood estimate (MLE) and (in this case) also the estimate obtained by ordinary least squares regression. It represents the parameter vector that yields the best possible fit to the observed data to be achieved by model M_k.

The combination of a Gaussian prior distribution and a Gaussian likelihood function yields an analytical expression for the posterior distribution $p(\mathbf{u}_k|\mathbf{y}_o, M_k) \sim \mathcal{N}(\tilde{\mathbf{u}}_k, \mathbf{C}_{\tilde{u}\tilde{u}})$, which is again Gaussian with $\tilde{\mathbf{u}}_k = \mathbf{C}_{\tilde{u}\tilde{u}} \left(\mathbf{C}_{\hat{u}\hat{u}}^{-1} \hat{\mathbf{u}}_k + \mathbf{C}_{uu}^{-1} \bar{\mathbf{u}}_k\right)$ and $\mathbf{C}_{\tilde{u}\tilde{u}} = \left(\mathbf{C}_{\hat{u}\hat{u}}^{-1} + \mathbf{C}_{uu}^{-1}\right)^{-1}$. Under the current set of assumptions, the mean of the posterior distribution, $\tilde{\mathbf{u}}_k$, is the maximum a posteriori estimate (MAP). The MAP represents those parameter values that are the most likely ones for model M_k, taking into account both prior belief about the distribution of the parameters and the performance in fitting the observed data. For a derivation of these statistics, see e.g., *Box and Tiao* [1973].

With the posterior parameter distribution, the quotient in equation (6) (Bayes' theorem rewritten to solve for the normalizing constant, equivalent to the integral in equation (7)) can be determined for any given value $\mathbf{u}_{k,i}$ within the limits of $p(\mathbf{u}_k)$.

3.2. Mathematical Approximation

If no analytical solution exists to the application at hand, equation (7) can be approximated mathematically, e.g., by a Taylor series expansion followed by a Laplace approximation. We briefly outline this approach in section 3.2.1 and then discuss the derivation of the KIC (section 3.2.2) which is based on this approximation. In this context, it becomes more evident how Occam's razor works in BMA (section 3.2.3). The BIC (section 3.2.4) represents a truncated version of the KIC. Another mathematical approximation, which is based on information theory, results in the AIC(c) (section 3.2.5). We contrast the expected impact of the different IC formulations on model selection in section 3.2.6.

3.2.1. Laplace Approximation

The idea of the Laplace method [*De Bruijn*, 1961] is to approximate the integral by defining a simpler mathematical function for a subinterval of the original parameter space, assuming that the contribution of this neighborhood almost makes up the whole integral. Here a Gaussian posterior distribution is assumed as simplification to the unknown distribution. This is a suitable approximation if the posterior distribution is highly peaked around its mode (or maximum) $\tilde{\mathbf{u}}$. This assumption holds, if a large data set with a high information content is available for calibration. Expanding the logarithm of the integrand in equation (7) by a Taylor series about the posterior mode $\tilde{\mathbf{u}}_k$ (i.e., the MAP), neglecting third-order and higher-order terms, taking the exponent again and finally performing the integration with the help of the Laplace approximation yields:

$$I_k = \mathcal{L}(M_k, \tilde{\mathbf{u}}_k|\mathbf{y}_o) p(\tilde{\mathbf{u}}_k|M_k) (2\pi)^{N_{p,k}/2} |\tilde{\Sigma}|^{1/2}, \tag{10}$$

with the likelihood function $\mathcal{L}(M_k, \tilde{\mathbf{u}}_k|\mathbf{y}_o)$, the prior density $p(\tilde{\mathbf{u}}_k|M_k)$, and the number of parameters $N_{p,k}$. The $N_{p,k}$ x $N_{p,k}$ matrix $\tilde{\Sigma} = -\left[\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u}^2}\right]^{-1}|_{\mathbf{u}=\tilde{\mathbf{u}}}$ is the negative inverse Hessian matrix of second derivatives and represents an asymptotic estimator of the posterior covariance $\mathbf{C}_{\tilde{u}\tilde{u}}$. It is equal to $\mathbf{C}_{\tilde{u}\tilde{u}}$ for the case of an

actually Gaussian posterior (see section 3.1). For details on the Laplace approximation in the field of Bayesian statistics and an analysis of its asymptotic errors, please refer to *Tierney and Kadane* [1986].

If the parameter prior is little informative, the expansion could also be carried out about the MLE $\hat{\mathbf{u}}_k$ instead of the MAP $\tilde{\mathbf{u}}_k$. This approximation will be less accurate in general, with the deterioration depending on the distance between the MAP and MLE estimators. However, the MLE may be easier to find than the MAP with standard optimization routines. The corresponding approximation takes the following form:

$$I_k = \mathcal{L}(M_k, \hat{\mathbf{u}}_k|\mathbf{y}_o)p(\hat{\mathbf{u}}_k|M_k)(2\pi)^{N_{p,k}/2}|\hat{\Sigma}|^{1/2}. \tag{11}$$

The inverse of the covariance matrix $\hat{\Sigma}$ is the observed Fisher information matrix evaluated at the MLE, $\mathbf{F} = -\frac{\partial^2 l}{\partial \mathbf{u}^2}|_{\mathbf{u}=\hat{\mathbf{u}}}$, with l being the log-likelihood function [*Kass and Raftery*, 1995].

If the normalized (per observation) Fisher information is used, $\mathbf{F}_1 = \mathbf{F}/N_s$, $\hat{\Sigma}$ equals $[\mathbf{F}_1 N_s]^{-1} = \frac{1}{N_s}\mathbf{F}_1^{-1}$ [*Ye et al.*, 2008]:

$$\begin{aligned} I_k &= \mathcal{L}(M_k, \hat{\mathbf{u}}_k|\mathbf{y}_o)p(\hat{\mathbf{u}}_k|M_k)(2\pi)^{N_{p,k}/2}|\hat{\mathbf{F}}|^{-1/2} \\ &= \mathcal{L}(M_k, \hat{\mathbf{u}}_k|\mathbf{y}_o)p(\hat{\mathbf{u}}_k|M_k)\left(\frac{2\pi}{N_s}\right)^{N_{p,k}/2}|\hat{\mathbf{F}}_1|^{-1/2}. \end{aligned} \tag{12}$$

For clarity in notation, model indices are omitted here for the covariance matrices and for the Fisher information matrix.

The presented mathematical approximations to the Bayesian Integral (equation (7)) are typically known in the shape of ICs, i.e., as $-2\ln I_k$. We subsequently discuss the three most commonly used ICs in the BMA framework. They all generally aim at identifying the optimal bias-variance trade-off in model selection, but differ in their theoretical derivation and therefore in their accuracy with respect to the theoretically optimal trade-off according to Bayes' theorem.

3.2.2. Kashyap Information Criterion

The Kashyap information criterion (KIC) directly results from the approximation defined in equation (12) by applying $-2\ln I_k$ [*Kashyap*, 1982]:

$$KIC_{\hat{u}} = -2\ln \mathcal{L}(M_k, \hat{\mathbf{u}}_k|\mathbf{y}_o) - 2\ln p(\hat{\mathbf{u}}_k|M_k) + N_{p,k}\ln\frac{N_s}{2\pi} + \ln|\hat{\mathbf{F}}_1|. \tag{13}$$

The KIC is applied within the framework of MLBMA [*Neuman*, 2002]. By means of this approximation, MLBMA is a computationally feasible alternative to full Bayesian model averaging if knowledge about the prior of parameters is vague. For applications of MLBMA, see *Neuman* [2003], *Ye et al.* [2004], *Neuman et al.* [2012], and references therein.

If an estimate of the postcalibration covariance matrix $\mathbf{C}_{\hat{u}\hat{u}}$ is obtainable, equation (11) can be drawn upon instead:

$$KIC_{\hat{u}} = -2\ln \mathcal{L}(M_k, \hat{\mathbf{u}}_k|\mathbf{y}_o) - 2\ln p(\hat{\mathbf{u}}_k|M_k) - N_{p,k}\ln(2\pi) - ln|\mathbf{C}_{\hat{u}\hat{u}}|. \tag{14}$$

Ye et al. [2004] point toward the close relationship of the $KIC_{\hat{u}}$ with the original Laplace approximation, but prefer the evaluation at the MLE, because it is in line with traditional MLE-based hydrological model selection and parameter estimation routines. *Neuman et al.* [2012] appreciate that MLBMA "admits but does not require prior information about the parameters" and include prior information in their likelihood optimization routine, which makes it de facto a MAP estimation routine. We strongly advertise the latter variant, because the Laplace approximation originally involves an expansion about the MAP instead of the MLE, and we understand prior information on the parameters as a vital part of Bayesian inference. Therefore, we propose to explicitly evaluate the KIC at the MAP:

$$KIC_{\tilde{u}} = \underbrace{-2\ln \mathcal{L}(M_k, \tilde{\mathbf{u}}_k|\mathbf{y}_o)}_{NLL} \underbrace{-2\ln p(\tilde{\mathbf{u}}_k|M_k)}_{1} \underbrace{-N_{p,k}\ln(2\pi)}_{2} \underbrace{-ln|\mathbf{C}_{\tilde{u}\tilde{u}}|}_{3}. \tag{15}$$

Occam factor

We will refer to this formulation as KIC@MAP as opposed to the KIC@MLE (equation (14)) to avoid any confusion within the MLBMA framework, which seems to admit both of the KIC variants discussed here. The

evaluation at the MAP is consistent with the Laplace approach to approximate the Bayesian integral and, in case of an actually Gaussian parameter posterior, will yield accurate results; this does not hold if the approximation is evaluated at the MLE. If the assumption of a Gaussian posterior is violated, it needs to be assessed how the different evaluation points affect the already inaccurate approximation. We will investigate the differences in performance between the KIC@MLE and our proposed KIC@MAP in section 4.

3.2.3. Interpretation Via Occam's Razor

In equation (15), we have distinguished different terms of the Laplace approximation in the formulation of the KIC@MAP. They can be interpreted within the context of BMA, and if the assumption of a Gaussian posterior parameter distribution is satisfied, this represents an interpretation of the ingredients of BME (or more specifically, minus twice the logarithm of BME). It incorporates a measure of goodness of fit (the negative log-likelihood term, NLL) and three penalty terms that account for model dimensionality (dimension of the model's parameter space). These three terms are referred to as Occam factor [*MacKay*, 1992].

The Occam factor reflects the principle of parsimony or Occam's razor: If any number of competing models shows the same quality of fit, the least complex one should be used to explain the observed effects. Any additional parameter is considered to be fitted to noise in the observed data and might lead to low parameter sensitivities and poor predictive performance (due to little robustness of the estimated parameters). Synthesizing the discussions by *Neuman* [2002, 2003], *Ye et al*. [2004], and *Lu et al*. [2011], and explicitly transferring them to the expansion about the MAP, we make an attempt to explain the role of the three terms that are contained in the Occam factor.

The parameter prior $p(\tilde{\mathbf{u}}_k|M_k)$ (term 1) implicitly penalizes a growing complexity in that it gives a lower probability density to models with larger parameter spaces (larger $N_{p,k}$), since high-dimensional densities have to dilute their total probability mass of unity within a larger space. Thus, a more complex model with its smaller prior parameter probabilities will obtain a higher value of the criterion or a decreased value of BME, which will compromise its chances to rule out its competitors according to Occam's razor.

The opposite is true for $-N_{p,k}\ln(2\pi)$ (term 2): here, an increase in dimensionality yields a decrease of the KIC or an increase in model evidence. This term is actually part of the normalizing factor of a Gaussian prior distribution and thus partially compensates the effect of (1).

Finally, $|\mathbf{C}_{\tilde{u}\tilde{u}}|$ (term 3) accounts for the curvature of the posterior distribution. A strong negative curvature, i.e., a very narrow posterior distribution, represents a high information content in the data with respect to the calibration of the parameters. A narrow posterior leads to a low value for the determinant and thus to a decrease in model evidence or an increase of the KIC. This might seem counter-intuitive at first, but has to be interpreted from the viewpoint that if the data provide a high information content, the resulting likelihood function shall be narrow, and thus, its peak value shall also be high. The determinant is thus a partially compensating counterpart to the NLL term. If two competing models achieve the same likelihood, but differ in their sensitivity to the data, the one with a smaller sensitivity will be chosen because of its robustness [*Ye et al*., 2010b].

3.2.4. Bayesian Information Criterion

The Bayesian information criterion (BIC) or Schwarz information criterion [*Schwarz*, 1978] is a simplification to equation (13) in that it only retains terms that vary with N_s:

$$BIC = -2\ln \mathcal{L}(M_k, \hat{\mathbf{u}}_k|\mathbf{y}_o) + N_{p,k}\ln N_s. \tag{16}$$

Evaluating this criterion for the MAP $\tilde{\mathbf{u}}_k$ would not be consistent, because the influence of the prior is completely ignored in equation (16). Since only *parts* of the Occam factor are retained compared to equation (15), the BIC penalizes a model's dimensionality to a different extent. Those differences are not supported by any specific theory. However, the KIC@MLE reduces asymptotically to BIC with growing data set size N_s. The reason is that the prior probability of $\hat{\mathbf{u}}_k$ as well as the normalized Fisher information do not grow with data set size, but the likelihood $-2\ln \mathcal{L}(M_k, \hat{\mathbf{u}}_k|\mathbf{y}_o)$ and $N_{p,k}\ln N_s$ do, rendering contributions that only grow with $N_{p,k}$ negligible [*Neuman*, 2003]. The error in approximation by the BIC is therefore expected to reduce to the error made by the KIC for large data set sizes. In section 3.2.6, we will compare the different IC approximations with regard to their penalty terms, and in section 4 we will investigate the convergence behavior of the KIC and BIC in more detail on a synthetic test case.

Applying the KIC or BIC for model *selection* (as opposed to averaging) is consistent as the assigned weight for the true model (if it is a member of the considered set of models) converges to unity for an infinite data set size. The truncated form (BIC) still seems to perform reasonably well for model identification or explanatory purposes [*Koehler and Murphree*, 1988]. It is also much less expensive to evaluate than the KIC for models with high-dimensional parameter spaces, since the evaluation of the covariance matrix is not required.

3.2.5. Akaike Information Criterion

The Akaike information criterion (AIC) or, as originally entitled, "an information criterion," originates from information theory (as opposed to the Bayesian origin of the KIC and BIC), but has frequently been applied in the framework of BMA [e.g., *Poeter and Anderson*, 2005]. It is derived from the Kullback-Leibler (KL) divergence that measures the loss of information when using an alternative model M_k with a predictive density function $g(Y|M_k, \mathbf{u}_k)$ instead of the "true" model with predictive density function $f(Y)$, with Y being a random variable from the true density f of the same size N_s as the observed data set $\mathbf{y}_o$:

$$\begin{aligned} D_{KL}(f,g) &= \int f(Y) \ln\left(\frac{f(Y)}{g(Y|M_k,\mathbf{u}_k)}\right) dY \\ &= E_f[\ln f(Y)] - E_f[\ln g(Y|M_k,\mathbf{u}_k)]. \end{aligned} \tag{17}$$

The first term in the second line of equation (17) is an unknown constant that drops out when comparing differences in the expected KL-information for the competing models in the set [e.g., *Kuha*, 2004]. *Akaike* [1973, 1974] argued that the second term, called relative expected KL-information, can be estimated using the MLE. For reasons not provided here, this estimator is biased by $N_{p,k}$, the number of parameters in the model M_k. Correcting this bias and multiplying by -2 yields the AIC [*Burnham and Anderson*, 2004]. The AIC formulation contains the NLL term as an expression for the goodness of fit and a penalty term for the number of parameters:

$$AIC = -2\ln \mathcal{L}(M_k, \hat{\mathbf{u}}_k|\mathbf{y}_o) + 2N_{p,k}. \tag{18}$$

Compared to the BIC in equation (16), the penalty term for the number of parameters $N_{p,k}$ is less severe. For data set sizes $N_s > 7$, BIC favors models with less parameters than AIC, since its penalty term $N_{p,k}\ln N_s$ becomes larger than the AIC's $2N_{p,k}$.

For a finite data set size N_s, a second-order bias correction has been suggested [*Sugiura*, 1978; *Hurvich and Tsai*, 1989]:

$$AICc = AIC + \frac{2N_{p,k}(N_{p,k}+1)}{N_s - N_{p,k} - 1}. \tag{19}$$

Among others, *Burnham and Anderson* [2004] suggest using the corrected formulation for data set sizes $N_s < 40N_{p,k}$. For increasing data set sizes, the AICc converges to the AIC.

The posterior Akaike model weight is derived from:

$$P(M_k|\mathbf{y}_o) = \frac{\exp(-0.5\triangle_k)}{\sum_{i=1}^{N_m}\exp(-0.5\triangle_i)} = \frac{\exp(-0.5AIC_k)}{\sum_{i=1}^{N_m}\exp(-0.5AIC_i)}, \tag{20}$$

with $\triangle_k = AIC_k - AIC_{min}$ or $\triangle_k = AICc_k - AICc_{min}$, respectively. Based on its theoretical derivation, the absolute value of AIC_k or $AICc_k$ has no explanatory power [*Burnham and Anderson*, 2004], only the difference $\triangle_k$ with respect to the lowest AIC_k or $AICc_k$ can be interpreted. The BIC or KIC, in contrast, are a direct approximation to BME (equation (7)) and therefore yield meaningful values, also in interpretation as absolute values.

The AIC seems to perform well for predictive purposes, with a tendency to over-fit observed data [see e.g., *Koehler and Murphree*, 1988; *Claeskens and Hjort*, 2008]. This tendency is supposedly less severe for the bias-corrected AICc. Both versions of the AIC do not converge to the true model for an infinite data set size. The reason is that, with an increasing amount of data, the model chosen by AIC(c) will increase in complexity, potentially beyond the complexity of the true model (if it exists) [*Burnham and Anderson*, 2004].

The KIC is expected to provide the most consistent results among the ICs investigated here because it is based on the approximation closest to the true equations. Applications and comparisons of KIC, BIC, and AIC can be found in *Ye et al.* [2008], *Tsai and Li* [2010], *Singh et al.* [2010], *Riva et al.* [2011], *Morales-Casique*

et al. [2010], and *Lu et al*. [2011]. In the following, we will summarize the main theoretical differences in the BME approximation by these ICs.

3.2.6. Theoretical Comparison of IC Approximations to BME

Based on equation (10), the Laplace approximated BME can be divided into the likelihood and the Occam factor *OF* (see section 3.2.3):

$$\check{I}_k = \mathcal{L}(M_k, \tilde{\mathbf{u}}_k|\mathbf{y}_o)p(\tilde{\mathbf{u}}_k|M_k)(2\pi)^{N_{p,k}/2}|\tilde{\Sigma}|^{1/2} = \mathcal{L}(M_k, \tilde{\mathbf{u}}_k|\mathbf{y}_o)OF. \tag{21}$$

The ICs analyzed here all share the same approximation for the goodness of fit term based on the MLE. The only exception is the KIC@MAP, which is evaluated at the MAP instead of at the MLE. However, they all differ in their approximation to the OF. The OF represents the penalty for the dimensionality of a model or what we call the *sharpness of Occam's razor*. For the different ICs, it is given by:

$$\begin{aligned} OF_{KIC,\tilde{u}} &= p(\tilde{\mathbf{u}}|M_k)(2\pi)^{N_{p,k}/2}|\mathbf{C}_{\tilde{u}\tilde{u}}|^{1/2} \\ OF_{KIC,\hat{u}} &= p(\hat{\mathbf{u}}|M_k)(2\pi)^{N_{p,k}/2}|\mathbf{C}_{\hat{u}\hat{u}}|^{1/2} \\ OF_{BIC} &= N_s^{-N_{p,k}/2} \\ OF_{AIC} &= \exp\left(-N_{p,k}\right) \\ OF_{AICc} &= \exp\left(-N_{p,k} - \frac{1}{2}\frac{2N_{p,k}\left(N_{p,k}+1\right)}{N_s - N_{p,k} - 1}\right). \end{aligned} \tag{22}$$

The OF as approximated by KIC does not explicitly account for data set size, yet N_s typically influences the curvature of the posterior probability (or the likelihood function) and thus implicitly affects $|\mathbf{C}_{\tilde{u}\tilde{u}}|^{1/2}$ (or $|\mathbf{C}_{\hat{u}\hat{u}}|^{1/2}$). In contrast, AICc and BIC explicitly take data set size N_s in account, but do not evaluate the sensitivity of the calibrated parameter set via the curvature. The effects of these differences on the accuracy of the BME approximation will be demonstrated exemplarily on two test case applications in sections 4 and 5.

3.3. Numerical Evaluation

Numerical evaluation offers a second alternative to determine BME, if no analytical solution is available or if one mistrusts the approximate character of the ICs. A comprehensive review of numerical methods to evaluate the Bayesian integral (equation (7)) is given by *Evans and Swartz* [1995]. In the following, we will shortly review selected state-of-the-art methods and discuss their strengths and limitations. Note that conventional efficient integration schemes (e.g., adaptive Gaussian quadrature) are limited to low-dimensional applications [*Kass and Raftery*, 1995]. In this study, we focus on numerical methods that can also be applied to highly complex models (models with large parameter spaces) in order to provide a useful discussion for a broad range of research fields and applications.

3.3.1. Monte Carlo Integration

Simple Monte Carlo integration [*Hammersley*, 1960] evaluates the integrand at randomly chosen points $\mathbf{u}_{k,i}$ in parameter space. These parameter sets $\mathbf{u}_{k,i}$ are randomly drawn from their prior distribution $p(\mathbf{u}_k|M_k)$. The integral (or expected value over parameter space, cf. equation (4)) is then determined as the mean value of the evaluated likelihoods (sometimes referred to as *arithmetic mean approach*):

$$\check{I}_k = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\left(M_k, \mathbf{u}_{k,i}|\mathbf{y}_o\right), \tag{23}$$

with the number of Monte Carlo (MC) realizations *N*. For large ensemble sizes *N* and a friendly overlap of the parameter prior and the likelihood function, this method will provide very accurate results. For high-dimensional parameter spaces, however, a sufficient (converging) ensemble might come at a high or even prohibitive computational cost. If the likelihood function is sharp compared to the prior distribution, only very few integration points will contribute a high likelihood value to the integral (making $\mathcal{L}$ a very skewed variable), and the numerical uncertainty in the approximated integral might be large.

3.3.2. Monte Carlo Integration With Importance Sampling

To reduce computational effort and improve convergence, importance sampling [*Hammersley et al*., 1965] aims at a more efficient sampling strategy. Instead of drawing random realizations from the prior parameter distribution, integration points are drawn from any arbitrary distribution that is more similar to the posterior

distribution. Thus, the mass of the integral will be detected more likely by the sampling points. When drawing from a different distribution q than the prior distribution p, the integrand in equation (7) must be expanded by q / q. This modifies equation (23) to:

$$\hat{I}_k = \frac{\sum_{i=1}^{N} w_i \mathcal{L}(M_k, \mathbf{u}_{k,i}|\mathbf{y}_o)}{\sum_{i=1}^{N} w_i}, \tag{24}$$

with weights $w_i = p(\mathbf{u}_{k,i}|M_k)/q(\mathbf{u}_{k,i}|M_k)$. The improvement achieved by importance sampling compared to simple MC integration will greatly depend on the choice of the importance function q.

3.3.3. Monte Carlo Integration With Posterior Sampling

Developing this idea further, it could be most advantageous to draw parameter realizations from the posterior distribution $p(\mathbf{u}_k|\mathbf{y}_o, M_k)$ in order to capture the full mass of the integral. Sampling from the posterior distribution is possible, e.g., with the MCMC method [*Hastings*, 1970].

For posterior sampling, the approximation to the integral reduces to the harmonic mean of likelihoods [*Newton and Raftery*, 1994], also referred to as the *harmonic mean approach*:

$$\hat{I}_k = \left[\frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(M_k, \mathbf{u}_{k,i}|\mathbf{y}_o)^{-1}\right]^{-1}. \tag{25}$$

Equation (25) can be subject to numerical instabilities, due to small likelihoods that may corrupt the evaluation of the harmonic mean.

According to Jensen's inequality [*Jensen*, 1906], sampling exclusively from the posterior parameter distribution yields a biased estimator that overestimates BME. Thus, the harmonic mean approach should be seen as a trade-off between accuracy and computational effort. In order to avoid the instabilities of the harmonic mean approach by solving equation (6) instead, it would be necessary to estimate the posterior parameter probability density from the generated ensemble through kernel density estimators [e.g., *Härdle*, 1991], which is only possible for low-dimensional parameter spaces. We therefore do not investigate this alternative option further within this study.

3.3.4. Integration in Likelihood Space With Nested Sampling

The main challenge in evaluating BME lies in sufficiently sampling high-dimensional parameter spaces. A promising approach which avoids this challenge and instead samples the one-dimensional likelihood space is called nested sampling [*Skilling*, 2006]. The integral to obtain BME is written as:

$$I_k = p(\mathbf{y}_o|M_k) = \int \mathcal{L}(M_k, \mathbf{u}_k|\mathbf{y}_o) dZ, \tag{26}$$

where Z represents prior mass $p(\mathbf{u}_k|M_k)du$. It is solved by discretizing into m likelihood threshold values with the sequence $\mathcal{L}_1 < \mathcal{L}_2 < \ldots < \mathcal{L}_m$ and summing up over the corresponding prior mass pieces $1 > Z_1 > Z_2 > \ldots > Z_m > 0$ according to a numerical integration rule. How to find subsequent likelihood thresholds is described in *Skilling* [2006].

One of the remaining challenges for this very recent approach lies in finding conforming samples above the current likelihood threshold. We follow *Elsheikh et al*. [2013] in utilizing a short random walk Markov chain starting from a random sample that overcame the previous threshold. However, instead of using the ratio of likelihoods as acceptance distribution, we take the ratio of prior probabilities instead, to ensure that new samples still conform with their prior. Another challenge lies in ending the procedure with a suitable termination criterion (e.g., stop if the increase in BME per iteration has flattened out or if the likelihood threshold cannot be overcome within a maximum number of MCMC steps).

If the prior mass enclosing a specific likelihood threshold was known, the value of BME could be determined as accurately as the integration scheme allows. However, the fact that the real prior mass pieces Z_j are unknown introduces a significant amount of uncertainty into the procedure, which reduces its precision. To quantify the resulting numerical uncertainty, an MC simulation over randomly chosen prior mass shrinkage factor $t_j = Z_j/Z_{j-1}$ should be performed [*Skilling*, 2006].

3.4. Conclusions From Theoretical Comparison

From our comparison of the underlying assumptions for the nine BME evaluation methods considered here, we conclude that out of the ICs, the KIC@MAP is the most consistent one with BMA theory. It represents the true solution if the assumptions of the Laplace approximation hold (i.e., if the posterior parameter distribution is Gaussian). The other ICs considered here represent simplifications of this approach or, in the case of the AIC(c), are derived from a different theoretical perspective and are therefore expected to show an inferior approximation quality. Among the numerical methods considered here, simple MC integration is the most generally applicable approach because it is bias-free and spares any assumptions on the shape of the parameter distribution, but is also the computationally most expensive one. The other numerical methods vary in their efficiency, but are expected to yield similarly accurate results, except for MC integration with posterior sampling, which yields a biased estimate. We will test these expectations on a synthetic setup in the following section. The underlying assumptions of the nine BME evaluation methods analyzed in this study are summarized in Table 4.

4. Benchmarking on a Synthetic Test Case

The nine methods to solve the Bayesian integral (equation (7)) differ in their accuracy and computational effort, as described in the previous section. To illustrate the differences in accuracy under completely controlled conditions, we apply the methods presented in section 3 to an oversimplified synthetic example. In a first step, we consider a setup with a linear model where an analytical solution exists. We create an ideal premise for the KIC@MAP and its variants (KIC@MLE and BIC) by using a Gaussian parameter posterior, which fulfills their core assumption. We designed this test case as a best-case scenario regarding the performance of these ICs: there is no less challenging case in which the information criteria could possibly perform better. In a second step, we also consider nonlinear models of different complexity that violate this core assumption. Since in this case no analytical solution exists, we use brute-force MC integration as reference for benchmarking. We designed this test case as an intermediate step toward real-world applications that typically entail nonlinear models and a higher number of parameters.

4.1. Setup and Implementation

In the first step, a linear model $\mathbf{y}=u_1\mathbf{x}+u_2$ relates bivariate Gaussian distributed parameters $\mathbf{u}=[u_1, u_2]$ (slope and intersect of a linear function) to multi-Gaussian distributed predictions $\mathbf{y}$ at measurement locations $\mathbf{x}$. This linear model is tested against a synthetic data set. The synthetic truth underlying the data set is generated from the same model, but with slightly different parameter values than the prior mean. To obtain a synthetic data set, a random measurement error is added according to a Gaussian distribution $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$.

For this artificial setup, BME can be determined analytically according to equations (9) (determining BME via linear uncertainty propagation) or 6 (solving Bayes' theorem for BME). This exact value is used as reference for the approximate methods presented in section 3. In this simple case, the MAP, $\tilde{\mathbf{u}}$, and the MLE, $\hat{\mathbf{u}}$, are known analytically. Also, the corresponding covariances $\mathbf{C}_{\tilde{u}\tilde{u}}$ and $\mathbf{C}_{\hat{u}\hat{u}}$ are known. We allow the mathematical approximations to take advantage of this knowledge by evaluating them at these exact values. Normally, these quantities have to be approximated by optimization algorithms first. This initial step represents the computational effort needed to determine BME with mathematical approximations in the form of ICs, since the evaluation of their algebraic equations itself is very cheap. Here we are not concerned with the potential challenge to find these parameter estimates, but merely wish to remind the reader of this fact.

The numerical evaluation schemes do not take advantage of the linearity of the test case. Their computational effort is determined by the number of required parameter realizations, and hence by the number of required model evaluations. To assess the expected improvement in accuracy by investing in computational effort, we repeat the determination of BME for increasing ensemble sizes (MC integration, importance sampling, sampling from the posterior) or increasing sizes of the active set (nested sampling). To determine the lowest reasonable ensemble size, we investigate the convergence of the BME approximation for each method. Simple MC integration is performed based on ensembles of 2000–1,000,000 parameter realizations drawn from the Gaussian prior. MC integration with importance sampling is performed based on the same ensemble sizes. The sampling distribution for importance sampling is chosen to be Gaussian with a mean value equal to the MAP and a variance equal to the prior variance. Realizations of the posterior parameter distribution for MC integration with posterior

sampling are generated by the differential evolution adaptive metropolis adaptive MCMC scheme DREAM [*Vrugt et al.*, 2008]. With DREAM, BME is approximated based on ensembles of 5000–1,000,000 parameter realizations. Convergence of the MCMC runs was monitored by the Gelman-Rubin criterion [*Gelman and Rubin*, 1992] and we chose to take the final 25% of the converged Markov chains as posterior ensemble. Nested sampling is performed with initial ensemble sizes of 10–10,000. A nested sampling run is complete if one of the two following termination criteria is reached. The first termination criterion stops the calculation if the current likelihood threshold could not be overcome within 100 MCMC steps with a scaling vector $\delta = 0.05\left[\sqrt{\mathbf{C}_{11}}, \sqrt{\mathbf{C}_{22}}\right]$. The second termination criterion stops the calculation if the current BME estimate would not increase by more than 0.5% even if the current maximum likelihood value would be multiplied with the total remaining prior mass. This results in total ensemble sizes (summed over all iterations) of roughly 1000–1,000,000.

As a base case, we generate a synthetic data set of size $N_s = 15$. Figure 1 shows the setup for the synthetic test case. The parameters used in the synthetic example are summarized in Table 1.

In our synthetic test case, the computational effort required for one model run is very low. This allows us to repeat the entire analysis for the base case and to average over 500 runs for each ensemble size in order to quantify the inherent numerical uncertainty in the results obtained from the numerical approximation methods. In the case of nested sampling, we additionally average over 200 random realizations of the prior mass shrinkage factor per run.

With the setup described above, we compare the performance of the different approximation methods in quantifying BME. Additionally, we study by scenario variations the impact of varied data set size and varied prior information (different mean values and variances of parameters) on the outcome of BME and on the performance of the different methods. The behavior of the mathematical approximations for small or large data set sizes has been touched upon in the literature [e.g., *Burnham and Anderson*, 2004; *Lu et al.*, 2011]. We will underpin these discussions by systematically increasing the data set size from $N_s = 2$ to $N_s = 50$. Again, the same model is used to generate the synthetic truth as in the base case, and the measurements are taken at equidistant locations on the same interval of $\mathbf{x}$. To show the general behavior of the approximation methods and to eliminate artifacts caused by a specific outcome of measurement error, we generate 200,000 perturbed data sets for each data set size and average over the results of these realizations.

To our knowledge, the impact of prior information on the performance of BME approximation methods has not yet been studied in such a systematic approach. With the help of our synthetic test case, we can assess and then discuss this impact in a rigorous manner. Figure 2a visualizes the prior parameter densities, the likelihood function, and the posterior densities for a range of prior widths, Figure 2b for different prior/likelihood overlaps. The second column represents the base case as described above. Variations in prior width are normalized as fractions of the base case variances (covariance is not varied), variations in overlap of prior and likelihood are measured as distance between the prior mean and the MLE. The varied parameter values are also listed in Table 1.

Besides the factors explained above, the model structure and dimensionality of the models' parameter spaces is expected to influence the performance of the different approximation methods. In the first step, we consider varying complexity with regard to the allowed parameter ranges as defined by the prior. In the second step, we also consider models with varied structure. Differences in model structure can manifest themselves in either differences in the dimensionality of the model (i.e., the number of parameters), or in the type of model (linear versus nonlinear), or in both. We consider all of these options here by including a linear model with one parameter $\mathbf{y} = u\mathbf{x}$ (smaller number of parameters, but same model type; L1), a weakly nonlinear model with two parameters $\mathbf{y} = \exp(u_1\mathbf{x}) + u_2$ (same number of parameters, but different model type; NL2), and a nonlinear model with four parameters $\mathbf{y} = u_1 \cos(u_2\mathbf{x} + u_3) + u_4$ (higher number of parameters and different model type; NL4) into the analysis. All prior distributions are chosen to be Gaussian with their mean and covariance values given in Table 1. For nonlinear models, no analytical solution exists. In order to still be able to assess the differences in approximation quality, we generate a reference solution with brute-force MC integration using a very large ensemble of 10 million realizations per model. We choose this exceptionally large number of realizations to obtain a very reliable estimate of BME as a reference. In a numerical convergence analysis (bootstrapping) [*Efron*, 1979], we determined the variance upon resampling of the ensemble members, which confirmed that the BME estimate is varying less than 0.001%. This variation is insignificant in relation to the lowest error produced by the compared BME evaluation methods, which is two orders of magnitude larger. The average BME approximation quality (and its

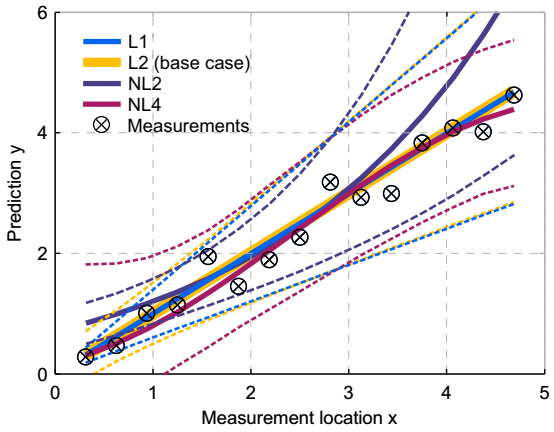


Figure 1. Synthetic test case setup. Measurements marked in black, prior estimate of linear (L1, L2) and nonlinear (NL2, NL4) models in solid lines, 95% Bayesian prediction confidence intervals in dashed lines of the respective color.

scattering) achieved by the other numerical methods is based on 500 repeated runs with ensemble sizes of 50,000, which might be considered a reasonable compromise between accuracy and computational effort based on the findings from the first step of our synthetic test case. From the posterior parameter sample generated by DREAM, we determine the MAP and the covariance matrix needed for the evaluation of the KIC@MAP. We obtain the respective ML statistics for the KIC@MLE from a DREAM run with uninformative prior distributions to cancel out the influence of the prior. We also evaluate the AIC(c) and the BIC at this parameter set.

For all cases (base case, varied data set size, varied prior information, varied model structure), the error in BME approximation is quantified as a relative error

$$E_{rel} = \frac{||\hat{I}_i - I||}{I}, \tag{27}$$

with the subscript i representing any of the discussed methods. In the case of numerical techniques, the average E_{rel} value and its Bayesian confidence interval out of all repetitions is provided.

Finally, we determine the impact of BME approximation errors on model weights based on the same setup and implementation details as described for the investigation of the influence of model structure (section 4.5).

4.2. Results for the Base Case

Figure 3 shows the relative error of BME approximations with respect to the analytical solution for the base case (see definition of parameters in section 4.1) as a function of ensemble size (number of model calls). Obviously, the accuracy of approximation improves for numerical methods when investing more computational effort, i.e., when increasing the numerical ensemble size. The improvement includes both a reduction

Table 1. Definition of Parameters Used in Different Scenarios of the Synthetic Test Case[a]

Parameter	Symbol	Value
Base Case (L2)		
Prior mean	$\bar{\mathbf{u}}$	$\bar{u}_1=1$; $\bar{u}_2=0$
Prior covariance	$\mathbf{C}_{uu}$	$C_{11}=C_{22}=0.04$; $C_{12}=C_{21}=-0.007$
Data set size	N_s	$N_s=15$
Meas. error covariance	$\mathbf{R}$	$R_{ii}=0.3^2$; $R_{ij}=0$
Varied Data Set Size		
Data set size	N_s	$N_s=2$–50
Varied Prior Width		
Prior covariance	$\mathbf{C}_{uu}$	$C_{11}=C_{22}=0.008$–2; $C_{12}=C_{21}=-0.007$
Varied Prior/Likelihood Overlap		
Prior mean	$\bar{\mathbf{u}}$	$\bar{u}_1=0.95$–1.5; $\bar{u}_2=-0.05$–0.5
Varied Model Structure		
Prior mean (L1)	$\bar{u}$	$\bar{u}=1$
Prior variance (L1)	s^2	$s^2=0.04$
Prior mean (NL2)	$\bar{\mathbf{u}}$	$\bar{u}_1=0.4$; $\bar{u}_2=-0.3$
Prior covariance (NL2)	$\mathbf{C}_{uu}$	$C_{11}=0.003$; $C_{22}=0.03$; $C_{12}=C_{21}=-0.0001$
Prior mean (NL4)	$\bar{\mathbf{u}}$	$\bar{u}_1=2.6$; $\bar{u}_2=0.5$; $\bar{u}_3=-2.8$; $\bar{u}_4=2.3$
Prior covariance (NL4)	$\mathbf{C}_{uu}$	$C_{11}=0.44$; $C_{22}=0.02$; $C_{33}=0.21$; $C_{44}=0.28$; $C_{12}=-0.07$; $C_{13}=0.24$; $C_{14}=-0.14$; $C_{23}=-0.05$; $C_{24}=0.02$; $C_{34}=-0.16$

[a]For variations of the base case, only differences to the base case parameters are listed.

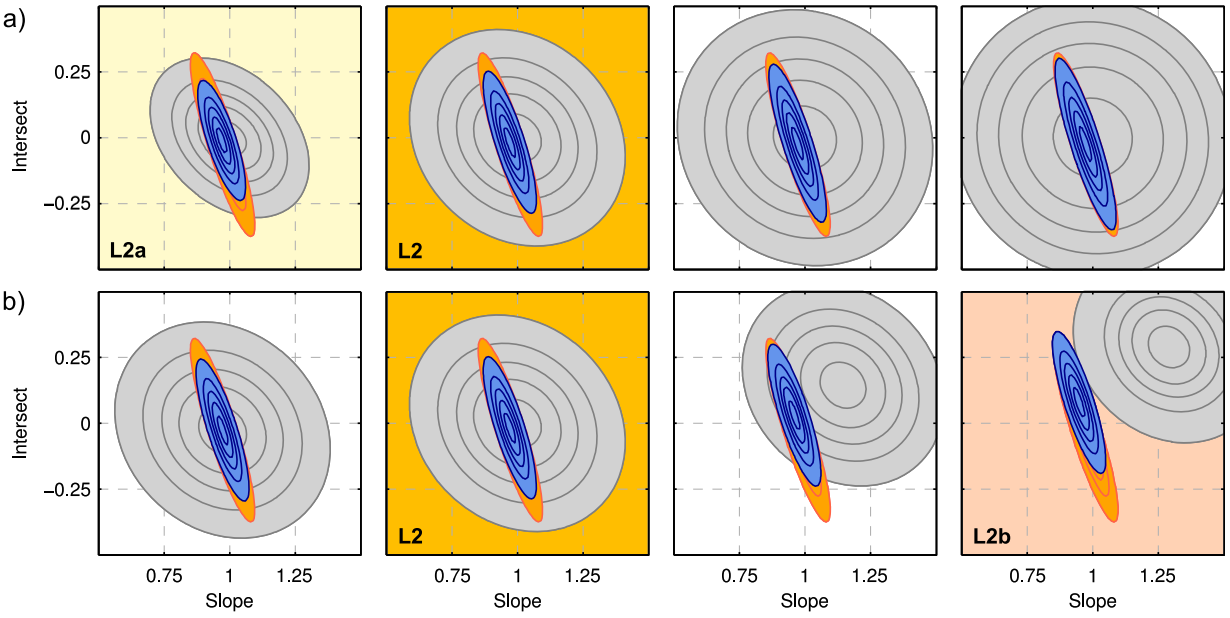


Figure 2. Prior densities (gray), likelihood (orange), and posterior densities (blue) for the different scenarios of the synthetic test case. Contour lines represent 10–90% Bayesian confidence intervals: (a) variations of prior width (fractions of base case variance shown here: 0.5,. . .,5), (b) variations of prior/likelihood overlap (distance between prior mean and MLE shown here: 0,. . .,0.3).

in bias (error) and a reduction in variance (numerical uncertainty, shown as 95% Bayesian confidence intervals of the approximation error in Figure 3).

Simple MC integration (MC) and MC integration with importance sampling (MC IS) perform equally well for this setup. MC results improve linearly in quality in this log-log-plot, which complies with its well-known convergence rate of $\mathcal{O}(N_s^{-1/2})$ (Central Limit Theorem) [*Feller*, 1968]. MC integration with sampling from the posterior (MC PS), however, leads to a severe overestimation of BME as anticipated (see section 3.3.3) and does not improve linearly with ensemble size in log-log-space, but shows a slower convergence. It also produces a much larger numerical uncertainty (keep in mind the logarithmic scale of the error axis). Note that the bias in BME approximation stems from the harmonic mean formulation and not from the sampling technique, because the posterior realizations generated with DREAM were checked to be consistent with the (in this case) known analytical posterior parameter distribution.

Nested sampling (NS) shows a similar approximation quality to MC integration, but is shifted on the *x* axis, i.e., it is less efficient with regard to numerical ensemble sizes in this specific test case. The convergence behavior shown here might not be a general property of nested sampling, because we found that modifications in the termination criteria significantly influence its approximation quality and uncertainty bounds. For this synthetic linear test case, we conclude that nested sampling is not as efficient as simple MC integration. It is also less reliable due to its somewhat arbitrary formulation with respect to the search for a replacement realization and the choice of termination criteria. In principle, it offers an alternative to simple MC integration and might become more advantageous in high-dimensional parameter spaces. We will continue this discussion for the real-word hydrological test case (section 5) and draw some final conclusions in section 6.

Since the ICs do not use random realizations to approximate BME, they are plotted as horizontal lines. With its assumptions fully satisfied, the KIC@MAP is equal to the analytical solution in this case. Therefore, it does not produce any error to be plotted in Figure 3. Evaluating the KIC at the MLE (KIC@MLE), however, leads to a significant deviation from the exact solution. For this specific setup, the AICc (after the KIC@MAP) performs best out of the mathematical approximations with a tolerable error of 3%. However, we will demonstrate later that this is not a general result. Note that we assess the AIC(c)'s performance in approximating the absolute BME value here for illustrative reasons, although strictly speaking, it is only derived for comparing models with each other, i.e., only the resulting model weights should be assessed (see section 4.6). The

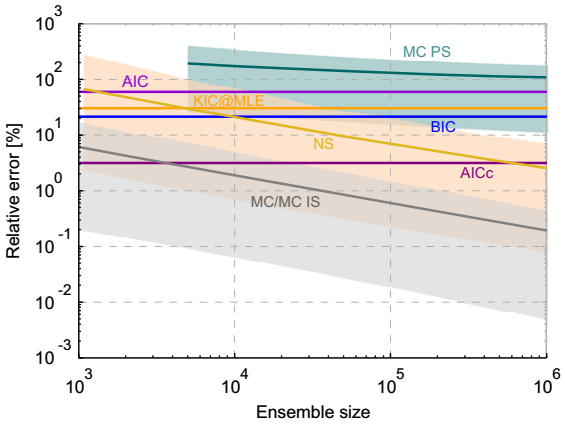


Figure 3. Relative error of BME approximation with respect to the analytical solution for the synthetic base case as a function of ensemble size. IC solutions are plotted as horizontal lines, as they do not use realizations for BME evaluation. Results of the numerical evaluation schemes are presented with 95% Bayesian confidence intervals.

other ICs yield approximation errors of 20–60%. Figure 3 shows that, except for MC integration with posterior sampling, the numerical methods outperform all of the ICs evaluated at the MLE, if only enough realizations are used.

4.3. Results for Varied Data Set Size

The approximation results as a function of data set size are shown in Figure 4. Since we have demonstrated that the numerical methods (except for posterior sampling) can approximate the true solution with arbitrary accuracy if only the invested computational power is large enough, we do not show their results here, as they would coincide with the solution of the KIC@MAP. Figure 4a shows the approximated BME values, while the relative error in percent with respect to the analytical solution is shown in Figure 4b.

The true BME curve (represented by the KIC@MAP here) is approximated quite well by both the KIC@MLE and by the BIC. However, while the KIC@MLE converges to the KIC@MAP with increasing data set size, the BIC does not. Its relative error with respect to the analytical solution becomes stable at more than 20%. This result is not in agreement with the findings of *Lu et al.* [2011], who confirmed the general belief that the BIC approaches the KIC with increasing data set size. In our case, the contribution of the terms dismissed by the BIC (see section 3.2) is still significant and hence produces a relevant deviation between the two BME approximations.

The AIC shows a linear dependence on data set size in this semilog plot. As expected, the AICc converges to the AIC with increasing data set size. Still, both variants of the AIC produce a relative error of more than 30%, which even increases with increasing data set size to more than 300% in this specific test case. Again, be reminded that the AIC(c) is only derived for comparing models with each other by the means of model weights, not as an approximation to the absolute BME value.

We investigate the reasons for the different behavior of the ICs over data set size by separating the likelihood term from the Occam factor penalty term (see section 3.2.6). Figure 4c shows how the true likelihood term (here: KIC@MAP) is approximated by the other ICs. Obviously, approximating this term produces negligible errors if the data set size is reasonably large, i.e., if the MAP and the MLE almost coincide. The problems in BME approximation clearly stem from the challenge of approximating the Occam factor (Figure 4d). The true Occam factor (or complexity penalty term) decreases with data set size. The KIC@MLE converges to this true behavior. The BIC is able to closely approximate the true curve, but does not yet converge to it in the range analyzed here. The penalty term of the AIC is a constant, which intersects the BIC's penalty term curve at $N_s = 7$ (see explanation in section 3.2.5). The penalty term of the AICc variant is converging to the constant AIC from below, i.e., it is increasing in contrast to the true, decreasing behavior. We conclude that the ICs differ substantially in the way they approximate the penalty term and therefore yield very different BME approximations with huge relative errors observed for the AIC and AICc.

Note that the results for $N_s = 15$ measurements, marked with circles in Figure 4, are similar, but not equal to the results we showed in Figure 3. This is due to the fact that to investigate the influence of data set size, we have marginalized over the random measurement error, while as a base case, we presented results for just one specific outcome of measurement error. We chose this scenario on purpose to illustrate that all those approximation methods which do not explicitly account for the sensitivity of the parameters to the specific data set, suffer from unpredictable behavior. The range of potential relative errors (95% Bayesian confidence intervals) over all 200,000 random realizations of measurement errors are shown as shaded areas in Figure 4b. It becomes clear that, up to a data set size of about $N_s = 20$ in our test case, none of the specific ICs would be a reliable choice: the AIC, AICc, and BIC could potentially yield very low (<1%) or very

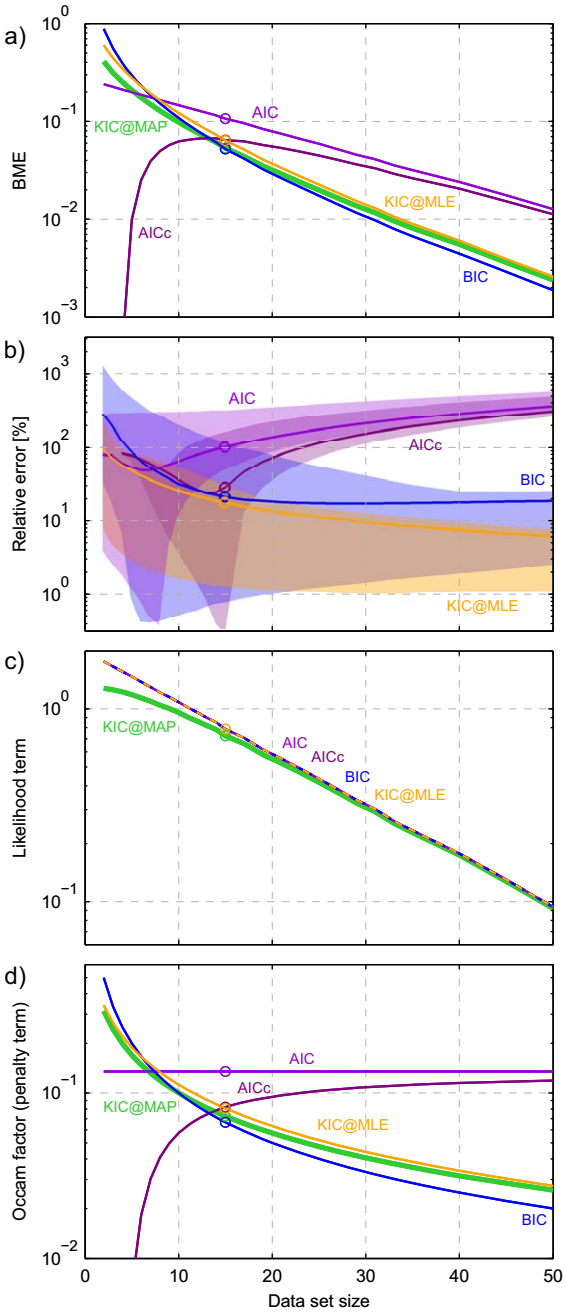


Figure 4. Synthetic test case results as a function of data set size: (a) approximation of BME, (b) relative error with respect to the analytical solution with 95% Bayesian confidence intervals, (c) likelihood term approximation, (d) Occam factor approximation. The result obtained from KIC@MAP represents the analytical solution in this case.

high (>100%) relative errors. Choosing the KIC@MLE is a more reliable choice, since it shows narrower bounds of potential relative errors, but still results in intolerable errors up to a data set size of about 30 in this case. We will elaborate on the question of how to make a safe choice of BME evaluation method in section 6.

4.4. Results for Varied Prior Information

Next, we investigate the behavior of the ICs for varied prior information. Figure 5 compares the influence of the prior width (left column) with the influence of varied distance between the MLE and the prior mean, i.e., of a shifted prior (right column). Again, we present the BME approximation (Figures 5a and 5e) and its relative error (Figures 5b and 5f), and the KIC@MAP represents the true solution. The AIC, AICc, and BIC approximations are constant over both variations, because they are not able to detect any information about the prior beyond the sheer number of parameters. In theory, however, increasing the prior width and moving the prior away from the area of high likelihood, both lead to a decrease in BME, which can be seen in the BME curve obtained by the KIC@MAP. While BME stabilizes at some point when increasing the prior width to a fully uninformative prior, it falls steeply if the prior is shifted farther away. This is important to keep in mind, because also the systematic relative errors in approximation (Figures 5b and 5f) are much larger for the shifted prior.

Only the KIC@MLE is able to track the variations in prior information and yields acceptable errors in both BME approximation and the approximation of the individual terms (likelihood term, Figures 5c and 5g, and penalty term, Figures 5d and 5h). Nevertheless, this error is in the range of 10%. In the case of increased prior width, the KIC@MLE converges to the true solution because the MAP moves toward the MLE, and, at the same time, the posterior covariance is approximated more closely by the covariance around the MLE. In contrast, there is no such convergence behavior with decreasing distance between the MLE and the prior mean, because in that case only the MAP moves toward the MLE, but the covariances do not coincide if the prior is still somewhat informative. Therefore, the solution of the KIC@MLE deviates from the true BME value even if the MAP is equal to the MLE, still producing a relative error of 30%.

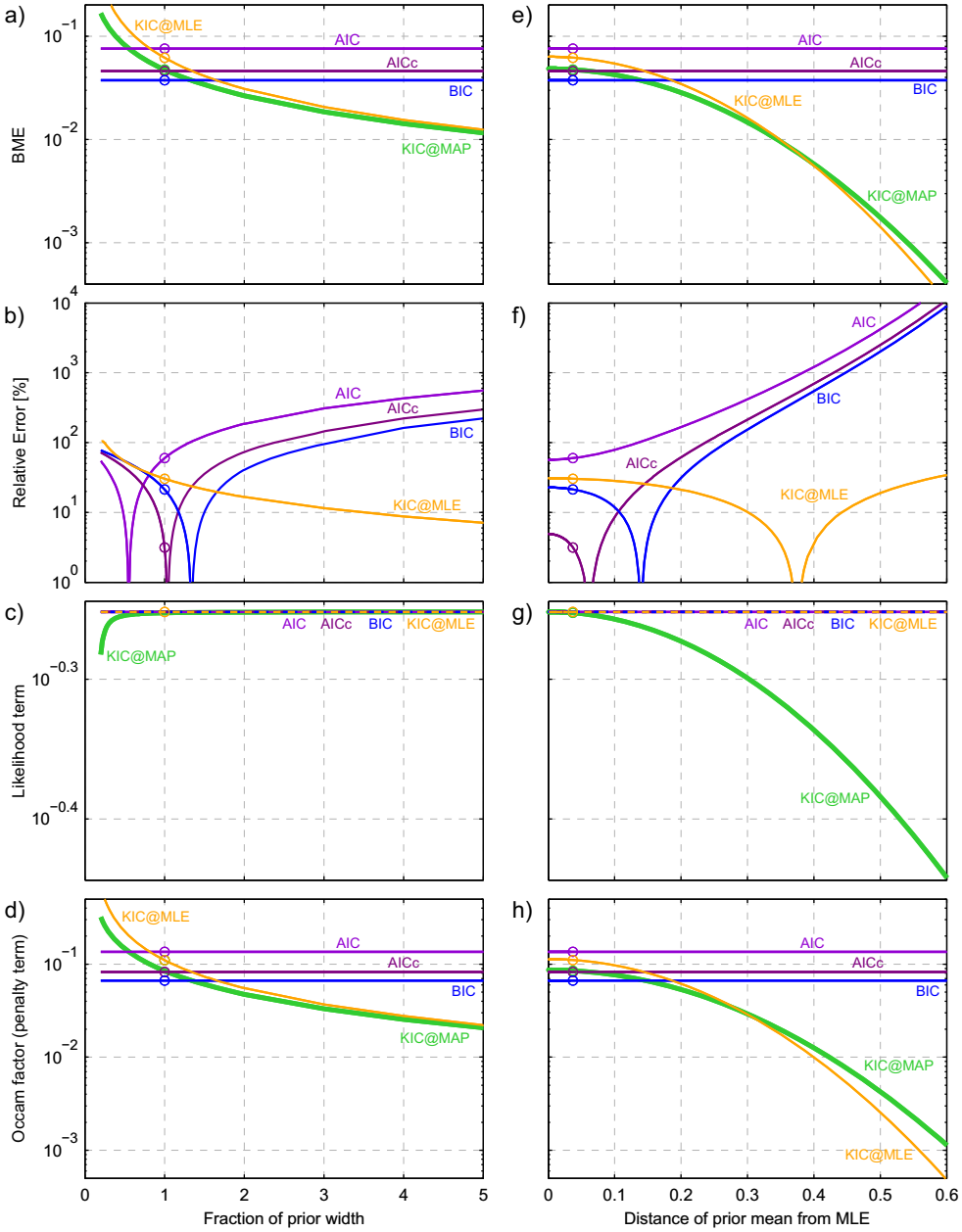


Figure 5. (left) Results obtained for the synthetic test case as a function of prior width and (right) as a function of distance between prior mean and MLE. (a and e) Approximation of BME, (b and f) relative error with respect to the analytical solution, (c and g) likelihood term approximation, (d and h) Occam factor approximation. The result obtained from KIC@MAP represents the analytical solution in this case.

For increasing distance between the MLE and the prior mean, the approximation of the likelihood term (Figure 5g) by MLE-based criteria deteriorates significantly. Since neither the likelihood term nor the penalty term are adequately approximated by the AIC, AICc, or BIC, substantial errors in BME approximation arise. There are poles in the relative error curves, where they cut the analytical solution. These locations are, however, dependent on the actual model at hand and on the outcome of the measurement error, and can therefore not be predicted a priori. Again, preferring any IC among AIC, AICc, and BIC as an approximation to BME is not a reliable choice as already pointed out when analyzing their performance over data set size. We will discuss implications of this finding in section 6.

4.5. Results for Varied Model Structure

In this section, we illustrate the influence of model structure on the BME approximation quality achieved by the nine different evaluation methods. In the previous sections, we have investigated the behavior of the ICs for varied data set size and varied prior information under optimal conditions, i.e., their underlying assumption of a Gaussian posterior distribution was fulfilled. For the nonlinear models considered here, this is no longer the case. This setup therefore represents a more realistic setting where no analytical solution exists. In order to still be able to assess the differences in approximation quality, we generate a reference solution with brute-force MC integration. We choose this method as reference for its absence of assumptions (section 3.3.1), i.e., its unrestricted applicability to any arbitrary (linear or nonlinear) setup, and for its precision and accuracy in BME approximation as demonstrated in the first step of our synthetic test case (section 4.2).

The relative approximation errors made by the different BME evaluation methods for the four different models are listed in Table 2. The performance of the numerical methods is comparable to the results shown in the previous sections, since their approximation quality is not directly related to model structure (but might be influenced by the shape of the area of high likelihood). Results for the AIC, AICc, and BIC vary arbitrarily with regard to model type and model dimensionality. We have shown that their approximation quality hugely depends on the actual data set (cf. section 4.3). This effect seems to be similarly strong here, mixing with errors due to the linear approximation of the complexity penalty term and due to violations of the underlying assumptions by the nonlinear models. The KIC variants show a much clearer tendency to fail with increasing nonlinearity of the model. The KIC@MAP is equal to the true solution in the case of the linear models L1 and L2, but not in the nonlinear case of the models NL2 and NL4. Since its approximation is perfect under linear (and multi-Gaussian) conditions, the deterioration in approximation quality for the nonlinear models clearly shows its deficiencies if these assumptions are not fulfilled. The KIC@MLE additionally suffers from differences in the location of the MAP and the MLE, which seems to cause similar trouble in the linear case (L2) and in the weakly nonlinear case (NL2). Note that the KIC@MLE suffers more strongly than the other ICs considered in this study, since not only its likelihood term, but also its Occam factor (penalty) term depends on this chosen point of expansion (see section 4.4).

4.6. Impact of Approximation Errors on Model Selection

The overestimation or underestimation of BME itself might not be a major concern, if it yielded consistent results in model weighting, i.e., if the estimated BME values were correlated with the exact values, so that ratios of BME between alternative models were consistent. Furthermore, the AIC(c) is derived to assess differences between competing models, and one would expect to see a better approximation to the true model weights than to the absolute BME values. To investigate this, we determine the model ranking for the four models described in section 4.1. We further introduce two additional versions of the base case model L2 by using two different prior distributions: Model L2a (see Figure 2) acts on an informative prior which has a significant overlap with the area of high likelihood. Model L2b uses a slightly less informative prior, which is significantly shifted away from the area of high likelihood. Model L2a is therefore clearly the favorite among those two model versions, because it makes better predictions while being even more parsimonious. We deliberately include those two versions as competing models to illustrate the inability of the AIC(c) and the BIC to detect differences in the parameter prior. We assign equal prior weights to all six models to let BME be the decisive factor in model averaging (see equation (2)).

Figure 6 shows the resulting posterior model weights as obtained from the different approximation methods, with the model weights of L2b, NL2, and NL4 additionally displayed on a logarithmic axis for better visual inspection. The ranking obtained from simple MC integration with an ensemble of 10 million realizations per model is used as reference solution. The solution obtained for the KIC@MAP coincides with the true weighting according to Bayes' theorem in case of the linear models L1, L2a, L2, and L2b, since the underlying assumption for the Laplace approximation is fulfilled. The results of simple MC integration (MC) and MC integration with importance sampling (MC IS) are shown in one bar in Figure 6 because, as demonstrated for the base case, they yield nearly identical results. MC integration with sampling from the posterior (MC PS) yields a biased result for model ranking, while nested sampling (NS) yields, on average, a very accurate result of model weighting, which indicates that the potential bias in overestimation or underestimation of BME induced by the somewhat arbitrary choice of termination criteria is consistent (correlated with the true value across the competing models).

Table 2. Relative Error of BME Approximation Methods for Different Model Structures as Compared to the Reference Solution (Analytical Solution Equal to the KIC@MAP in Case of Linear Models, Brute-Force MC Integration in Case of Nonlinear Models, Highlighted in Italic Font)[a]

Method	$E_{rel,L1}$ (%)	$E_{rel,L2}$ (%)	$E_{rel,NL2}$ (%)	$E_{rel,NL4}$ (%)
KIC@MLE	0.9	30.4	24.9	99.8
KIC@MAP	*0.0*	*0.0*	13.2	59.4
BIC	94.0	21.3	37.5	70.3
AIC	176.4	59.7	179.2	22.5
AICc	137.0	3.2	69.4	83.4
MC	0.0	0.0	*0.0*	*0.0*
MC IS	0.8 [0.0; 2.2]	0.1 [0.0; 0.4]	1.1 [0.0; 2.9]	1.5 [0.1; 4.1]
MC PS	132.0 [25.7; 196.6]	131.8 [23.5; 221.7]	324.8 [40.8; 490.3]	232.8 [18.1; 481.8]
NS	2.4 [0.1; 6.4]	2.8 [0.4; 27.2]	4.0 [0.2; 10.7]	11.2 [3.3; 18.9]

[a]95% Bayesian confidence intervals of numerical results given in parentheses.

The AIC, AICc, and BIC assign a too large weight to the simplest model in the set (L1). This is due to the fact that these criteria merely count the number of parameters used by a model, instead of considering correlations among the parameters (defined by the parameter prior) which reduce the actual degrees of freedom. Furthermore, these criteria are not able to distinguish between the three models L2a, L2, and L2b, which only differ in their prior parameter assumptions, but not in their respective MLE. The corresponding BME approximations by AIC, AICc, and BIC therefore yield the most indecisive weighting for these three models (equal weights), whereas the true BMA weights convey the clear message that, out of these three models, L2a should be preferred over L2, and L2b should be discarded.

Within this model set, the nonlinear models obtain very small weights (see Figure 6) in the reference solution because the model structure of NL2 does not match the data well and NL4 is already too complex to compete with the simpler linear models. Since the BME approximation errors of the KIC variants increase drastically with the nonlinearity of the models, these errors are expected to impact model ranking significantly if nonlinear models are playing a relevant role in the model selection competition. Here the nonlinear models play an almost irrelevant role and thus model ranking is not too badly compromised when using the KIC@MLE or the KIC@MAP, with the latter still outperforming the former.

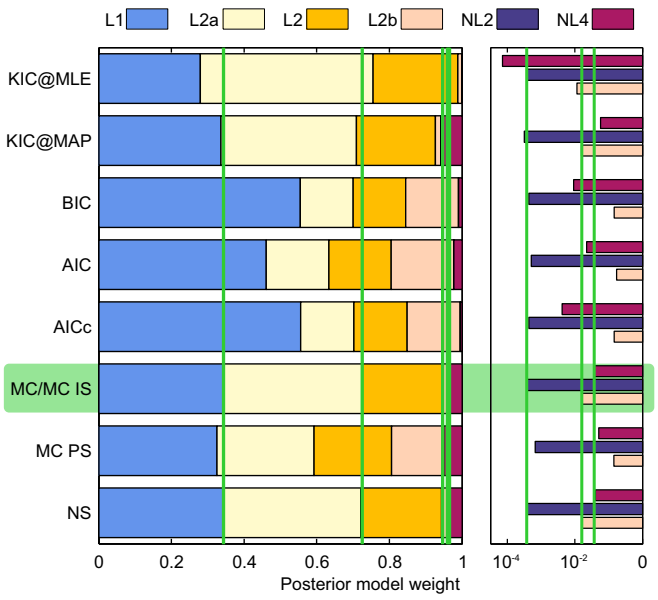


Figure 6. Posterior model weights as obtained from the different BME evaluation methods for linear (L1, L2) and nonlinear (NL2, NL4) models. L2, L2a, and L2b represent the same linear model with differently shaped priors. Green vertical lines indicate reference solution (obtained from brute-force MC integration).

4.7. Conclusions From Synthetic Test Case

The benchmarking has shown that all ICs (except for the KIC@MAP) potentially yield unacceptably large errors in BME approximation. Their performance depends on the actual data set (including the outcome of measurement error) that is used for calibration. We have learned that the AIC and AICc behave differently from the BIC and KIC for increasing data set size, i.e., the error made by the AIC(c) increases, while the error of the BIC and KIC decreases. Under varied prior information, however, the BIC follows the error behavior of the AIC(c) in that it cannot distinguish models which only differ in their prior definition of the parameter space. This is a crucial finding, since the prior contains all the information about the

flexibility of a model and is the basis for an adequate punishment of model complexity. Also, differences in model dimensionality cannot be adequately captured by those ICs. Among the ICs considered here, the KIC@MLE is closest to the true solution, which is identical to the KIC@MAP in the linear case. The performance of the KIC under nonlinear conditions deteriorates toward unacceptable approximation errors in both its versions, while the KIC@MAP still outperforms the KIC@MLE. Except for MC integration with posterior sampling, we have shown that the numerical methods considered here are capable of approximating the true BME value with satisfying accuracy, if the required computational power is affordable. Out of the numerical methods, simple MC integration and MC integration with importance sampling show the highest accuracy and lowest uncertainty for a given computational effort, which is in agreement with the theoretical basis of the respective methods (see section 3.4). Our findings regarding the approximation quality of the absolute BME value also apply to the approximation quality of the resulting model weights in this synthetic case, with one exception: nested sampling proves to yield a similarly accurate model ranking despite its slightly higher BME approximation errors as compared to MC integration.

5. Real-World Application to Hydrological Model Selection

In this section, we describe the application of the BME approximation methods presented in section 3 to real-world hydrological model selection. Due to the nonlinearity in the hydrological models considered here, no analytical solution exists. As already argued in section 4.5, we generate a reference solution by investing a large amount of computational effort into a brute-force MC integration. Hydrological model selection based on discharge measurements as presented here can be seen as a relatively simple model selection task, since generally a large number of measurements is available, which emphasizes differences in model behavior. In other disciplines, model selection might become more difficult, as the number of measurements and their information content are typically limited. Therefore, this real-world application can still be considered as a rather good-natured case of model selection. BME evaluation methods which fail in this application are expected to perform similarly insufficiently or even worse in other applications.

5.1. Setup and Implementation

We use the distributed mesoscale hydrologic model (mHM, Version 4.0) [*Samaniego et al*., 2010] to illustrate the performance of the different BME approximation methods in hydrological model selection. mHM is based on numerical approximations of dominant hydrological processes that have been tested in various existing hydrological models (e.g., in HBV [*Bergström et al*., 1997] and VIC [*Liang et al*., 1996]). It features a novel multiscale parameter regionalization technique to treat subgrid variability of input variables and model parameters [*Samaniego et al*., 2010; *Kumar et al*., 2010]. A detailed description of mHM can be found in *Samaniego et al*. [2010], *Kumar et al*. [2010], and *Wöhling et al*. [2013] and is therefore not repeated here. The model is applied to the Fils river catchment (area 361 km^2) of the Upper Neckar basin, Southwest Germany, using daily discharge measurements for the time period between 1980 and 1988. Please refer to sub-catchment 17 in *Wöhling et al*. [2013] for details on the model setup and a multicriteria model calibration. The original model considers two soil layers and employs 53 global parameters, 33 of which have been found sensitive to discharge predictions in a sensitivity analysis conducted prior to this study (results not shown here). This model is subsequently referred to as mHM2L. For the purpose of this study, a slightly simpler model with a single soil layer was built (mHM1L), where 29 of the 53 global parameters have been found sensitive to discharge predictions. Conventional model calibration for these two models yields Nash-Sutcliffe efficiencies (NSE) [*Nash and Sutcliffe*, 1970] of 0.9309 and 0.9073 for mHM2L and mHM1L, respectively. These values indicate that both models are able to adequately reproduce the observed discharge time series.

A uniform prior is assumed for the sensitive parameters with parameter ranges set to mHM-recommended values [see *Kumar et al*., 2010]. The insensitive parameters are fixed at midrange.

From a preprocessing analysis, the residuals between predictions and observations were found to be heteroscedastic, which is often encountered in hydrological modeling [*Sorooshian and Dracup*, 1980]. To mitigate heteroscedasticity, we applied a Box-Cox transformation [*Box and Cox*, 1964] to both discharge predictions $\mathbf{y}_i$ and discharge observations $\mathbf{y}_o$:

$$\mathbf{y}' = \frac{\mathbf{y}^{\lambda} - 1}{\lambda}. \tag{28}$$

A value of $\lambda = 0.55$ proved to be best suitable to reduce heteroscedasticity and to achieve a satisfying compromise in the fit to both high-flow and low-flow periods. The remaining variance in residuals is attributed to both measurement noise and conceptual errors. If only tested against measurement noise, both models (just like any other conceptual model) would be rejected from the statistical viewpoint [*Reichert and Mieleitner*, 2009] despite the high NSE values reported above, because not all observations can be reproduced within measurement error bounds. The residuals appeared to be correlated to a varying extent on different time scales (most likely due to superposition of seasonal trends and event-based model deficiencies), such that we could not identify a parsimonious correlation model that would reasonably explain the observed patterns. Since the identification of more elaborate but robust error models [see e.g., *Del Giudice et al*., 2013; *Evin et al*., 2014] is beyond the scope of our study, we restricted ourselves to a simple uncorrelated error that increases with discharge in the untransformed space, similar to a relative error. We chose a relatively large and constant standard deviation of 4.4 $\left(\mathrm{m^3 s^{-1}}\right)^{0.55}$ for this lumped error in transformed space as a mild penalty for deviations from the observations. The sheer length of the observed time series nevertheless led to an effective reduction of parameter and prediction uncertainty, i.e., to a successful and reliable calibration. Posterior uncertainty in discharge predictions was reduced to around 5% of prior uncertainty for both models. Our chosen error parameterization is represented as an error matrix $\mathbf{R}'$ with entries on the main diagonal. The Gaussian likelihood function to be evaluated in equation (7) then takes the form:

$$p\left(\mathbf{y}_o | M_k, \mathbf{u}_{k,i}\right) = 2\pi^{-N_s/2} |\mathbf{R}'|^{-1/2} \cdot \exp\left(-\frac{1}{2}\left(\mathbf{y}_i' - \mathbf{y}_o'\right)^T \mathbf{R}'^{-1}\left(\mathbf{y}_i' - \mathbf{y}_o'\right)\right) \prod_{j=1}^{N_s} \mathbf{y}_{o,j}^{\lambda - 1}, \tag{29}$$

with the last term being the derivative of the transformed observations $\mathbf{y}_o'$ with respect to the untransformed observations $\mathbf{y}_o$. Prior and posterior predictions of discharge as obtained from both models are shown for the first year of the observation time series in Figure 7.

The reference BME value for each model was determined by MC integration (equation (23)) over ensembles of 1.1 million realizations per model. Effective sample sizes [*Liu*, 2008] (the number of prior realizations that significantly contribute to the posterior distribution) of 95 for the double-layer model (mHM2L) and of 27 for the single-layer model (mHM1L) were considered to be sufficiently large to produce a reliable statistic of BME. We again performed a numerical convergence analysis (see section 4.1), which confirmed that the BME estimate is varying less than 1%.

Based on this reference solution, we assess the approximation quality of the different methods presented in section 3. From a posterior parameter ensemble of size 25,000 generated by DREAM, we determine the MAP (and the covariance matrix needed for the evaluation of the KIC). In this case, the MAP is equal to the MLE due to the uniform prior parameter distributions, and because the MLE does not fall outside of the bounds of the prior. We evaluate the AIC, BIC, and KIC at this parameter set. Note that, when using the full calibration time series with $N_s = 3227$ measurements, we do not consider the AICc, because it yields the same results as the AIC. This is expected since the ratio N_s/N_p is much larger than 40 and, therefore, there is no need to correct for a small data set size (see section 3.2.5). Importance sampling is performed with Gaussian sampling distributions that are centered about the MLE (obtained from DREAM) with the same standard deviations as specified for the prior parameter distributions. 100,000 realizations are generated from each model. Nested sampling is performed with an initial ensemble size of 500. The same termination criteria as described in the synthetic test case (section 4) are used. This leads to a total ensemble size (number of model calls) of around 110,000.

All BME approximation methods are assessed with regard to the error in reproducing the absolute BME value as well as with regard to model ranking. Since all mathematical approximation methods drastically underestimate BME in this test case and relative errors would all end up being 100%, we specify the approximation error instead by determining the ratio between approximate and reference BME value and taking the natural logarithm:

$$E_{ln(r)} = \ln \frac{\tilde{I}_i}{I} = \ln \tilde{I}_i - \ln I. \tag{30}$$

A good approximation will hence be characterized by a ratio of close to 1 and an $E_{ln(r)}$ value of close to 0. A negative value indicates that the approximation underestimates the reference value, and a positive value

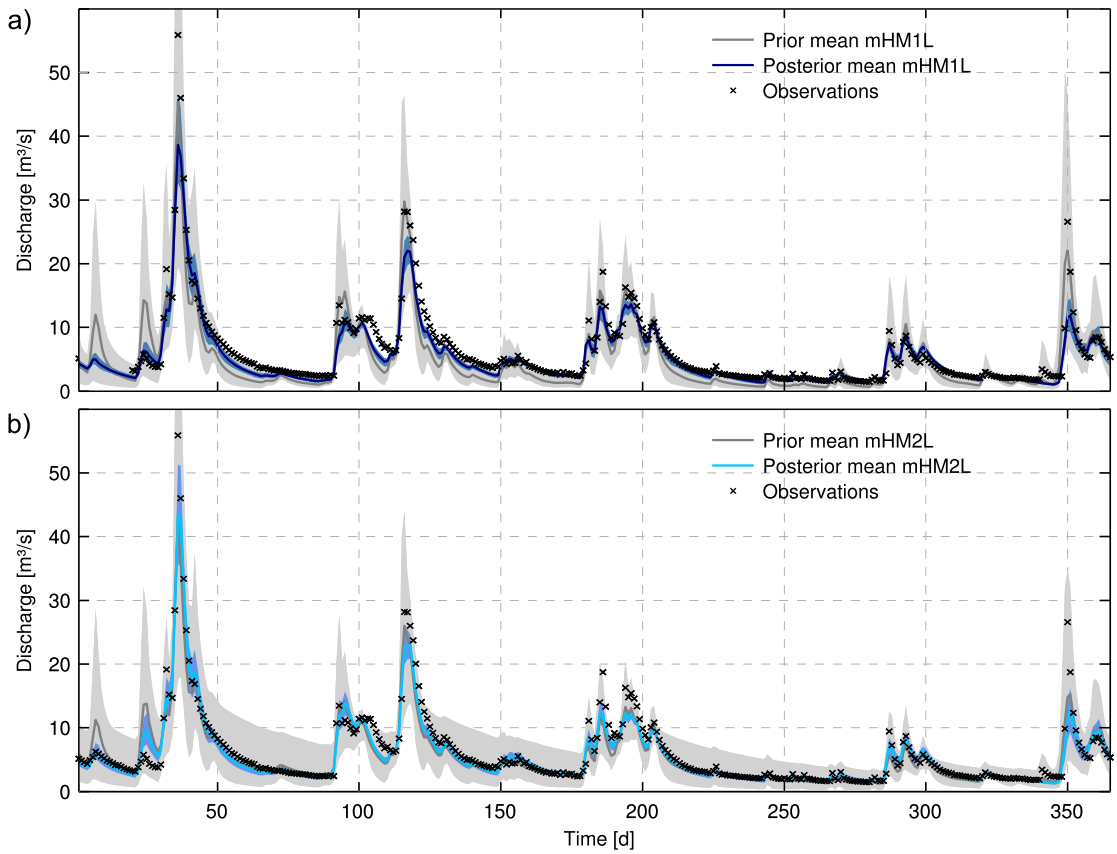


Figure 7. Model predictions of discharge for first year of calibration time series. (a) mHM1L, (b) mHM2L. Prior 95% Bayesian confidence intervals are shaded in gray, posterior 95% Bayesian confidence intervals in respective color.

indicates overestimation. The impact on model ranking is evaluated by comparing the resulting model weights (equation (2)).

5.2. Results for Full Observation Time Series

When using the full observation time series of 3227 daily discharge measurements for calibration, BME takes a very small value since the likelihood of a parameter set is the product of the relatively small likelihoods for each data point. We therefore report the natural logarithm of the BME value instead. The reference ln(BME) values as obtained from MC integration are −10,003 for the single-layer model (mHM1L) and −9991 for the double-layer model (mHM2L). While the numerical methods and the KIC arrive relatively close to this reference solution, the AIC and BIC drastically underestimate the reference values (see Table 3). Again, the AIC might be excused because it is not derived for approximating the absolute BME value, but for approximating the differences in models via model weights. The numerical methods tend to overestimate BME (especially posterior sampling and importance sampling), while the KIC underestimates instead. Nested sampling yields the most accurate results of all numerical methods (besides the MC reference solution that has a roughly 10-fold computational effort). MC integration with importance sampling leads to an overestimation of BME here. This indicates that the choice of importance density is already close enough to the posterior parameter density to inherit (some of) the biasedness of MC integration with posterior sampling (see section 3.3.3). Results of all approximation methods are summarized in Table 3.

Based on BME values, the double-layer model mHM2L is the clear winner in this model comparison. Although both models exhibit a good predictive performance which is confirmed by NSE values greater than 0.90, mHM2L outperforms mHM1L because the concept of two soil layers is able to mimic the water

Table 3. Performance of BME Approximation Methods in Hydrological Test Case as Compared to the Brute-Force MC Reference Solution (Highlighted in Italic Font)

	Full Time Series		Reduced Time Series	
Method	$E_{ln(r),mHM1L}$	$E_{ln(r),mHM2L}$	$E_{ln(r),mHM1L}$	$E_{ln(r),mHM2L}$
KIC	−5.2	−7.2	−6.8	−8.4
BIC	−103.1	−120.6	−79.7	−92.1
AIC	−15.0	−20.3	−23.1	−27.8
AICc			−25.7	−31.2
MC	*0.0*	*0.0*	*0.0*	*0.0*
MC IS	1.4	3.3	1.0	1.5
MC PS	3.1	4.8	0.8	1.0
NS	1.1	0.5	0.3	−0.1

retention capacity and the water movement in the unsaturated zone better than a single soil layer. This leads to a slightly but notably better fit of the simulated discharge hydrograph to the data. With regard to model complexity, both models use the same number of parameters, but differ in the number of sensitive (nonfixed) parameters. Even though the number of sensitive parameters is slightly smaller for mHM1L than mHM2L (29 versus 33), the size of the parameter space as defined by the prior density is larger in the case of mHM1L due to the actual permitted range of parameter values; i.e., some sensitive parameters of mHM1L, which are fixed in mHM2L, have a larger range of values. As a consequence, there is no trade-off between goodness of fit and model complexity, and model mHM2L obtains a full posterior model weight of 100% (mHM1L obtains a weight of 0.0004% in the reference solution).

All numerical methods as well as the KIC and AIC agree with the reference model weights. The BIC, however, yields the exact opposite: It assigns a weight of 99.4% to the competing model, mHM1L. The BIC's parameterization of the Occam factor is in this case the decisive (and guilty) factor, because the difference in the countable number of parameters is large enough to compensate the slightly better goodness of fit, which is in reality not the case. Posterior model weights are visualized in Figure 8a. Although the quality of approximation achieved by the different methods varies drastically, the model ranking and model selection result is (except for BIC) extremely clear and correct. This is due to the fact that the difference in goodness of fit is significantly large, in favor of mHM2L. This is mirrored in the difference of maximum likelihoods between both models, and is therefore detected by the ICs.

Previous studies on hydrological model selection have yielded similar results in that one model obtained an IC weight of close to 100% [e.g., *Meyer et al*., 2007]. *Lu et al.* [2013] have interpreted this clear weighting as too "aggressive," because it did not seem justified given the available data and prior knowledge. They therefore propose to use a different formulation for the likelihood function which considers not only uncorrelated (measurement) errors, but also correlated model structural errors. This led to a less clear weighting for all of the ICs that are also considered here. *Tsai and Li* [2010] proposed a different procedure, but following the same line of thought: They suggest to modify the calculation of the model weights by scaling the IC results in order to obtain a less obvious model ranking. *Ye et al*. [2010a] used the generalized likelihood uncertainty estimation method (GLUE) [*Beven and Binley*, 1992]), which penalizes deviations from the observed values in a nonformal, user-specified manner. They calculated posterior model weights merely based on GLUE-likelihoods of the calibrated models (which does not incorporate any Bayesian penalty for complexity) and compared them with model weights obtained from ICs, also finding that ICs seem to discriminate too harshly. *Rojas et al.* [2008], on the other hand, compared different (formal and informal) likelihood formulations within the theoretical BMA framework, but did not find significant differences in the resulting model weights.

We would like to add to this discussion by pointing out that the choice of likelihood function clearly influences the outcome of BME and, potentially, the resulting model weights as a consequence. If uncorrelated errors are assumed, the likelihood for all individual data points are multiplied. If, for a long data series, one model at all times predicted only slightly worse than its opponent, multiplying all the slightly lower likelihoods would lead to a drastically lower total likelihood and hence to a significantly lower BME value. BMA theory therefore supports the finding that, if only enough data points are available for calibration (and such a type of likelihood function is used), one model will be the clear winner—no matter if the competing model is performing very well, too, as it is the case in our application.

This might seem counter-intuitive, as one would argue that there is not necessarily more information about model choice in a more frequently measured time series. But it needs to be kept in mind that the clear model ranking is only based on the calibration period, i.e., the models might behave very differently under changing boundary conditions. Hence, the choice of the calibration period, the sampling frequency and the definition of the likelihood function lie in the responsibility of the modeler and should be subject to further

Water Resources Research

10.1002/2014WR016062

Table 4. Overview of Methods to Evaluate Bayesian Model Evidence

Evaluation method	Abbreviation	Eq.	Underlying Assumptions	Comp. Effort	Performance in Linear Test Case	Performance in Non-linear Test Cases	Recommended Use
Analytical solution							
Theoretical distribution of BME	-	9	Gaussian parameter prior and likelihood, linear model	Negligible	Exact	Not available	Whenever available
Normalizing constant of parameter posterior	-	6	conjugate prior, linear model	Negligible	Exact	Not available	Whenever available
Mathematical approximations							
Kashyap's information criterion, evaluated at MLE	KIC@MLE	14	Gaussian parameter posterior, negligible influence of prior	Medium	Relatively accurate (assumptions mildly violated)	Inaccurate	KIC@MAP to be preferred
Kashyap's information criterion, evaluated at MAP	KIC@MAP	15	Gaussian parameter posterior	Medium	Exact (assumptions fulfilled)	Inaccurate	If assumptions fulfilled/ numerical techniques too expensive
Bayesian information criterion	BIC	16	Gaussian parameter posterior, negligible influence of prior	Low	Potentially very inaccurate (depending on actual data set), ignores prior		Not recommended for BMA
Akaike information criterion	AIC	18	(not derived as approximation to BME)	Low	Potentially very inaccurate (depending on actual data set), ignores prior		Not recommended for BMA
corrected Akaike information criterion	AICc	19	(not derived as approximation to BME)	Low	Potentially very inaccurate (depending on actual data set), ignores prior		Not recommended for BMA
Numerical evaluation techniques							
Simple Monte Carlo integration	MC	23	None	Extreme	Slow convergence, but bias-free		Whenever computationally feasible
MC integration with importance sampling	MC IS	24	None	High	Faster convergence, but (potentially) biased		As a more efficient alternative to MC
MC integration with posterior sampling	MC PS	25	None	High	Even faster convergence, but even more biased (due to harmonic mean approach)		Not recommended for BMA
Nested sampling	NS	26	None	High	Slow convergence for BME (due to uncertainty in prior mass shrinkage), but bias-free		Promising alternative to MC, more research needed

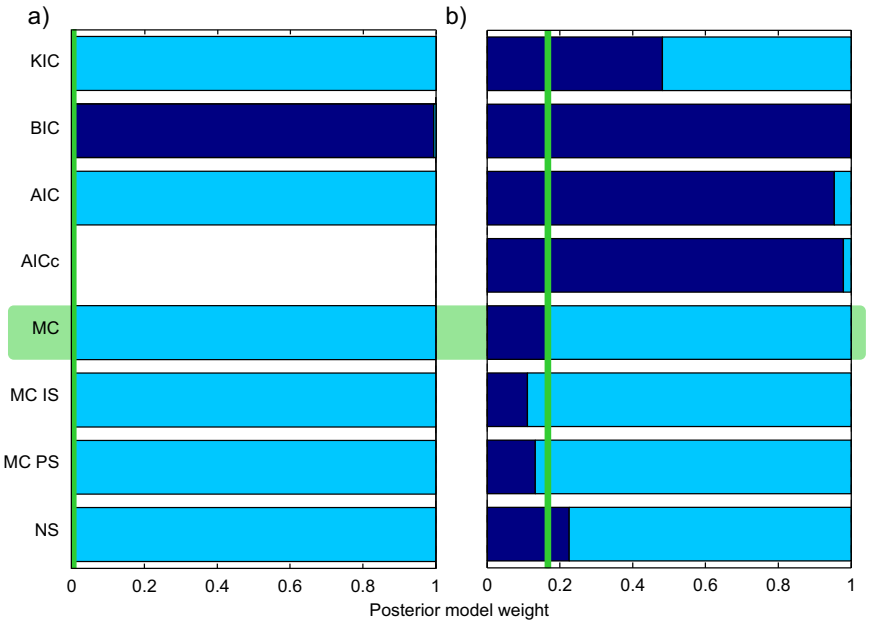


Figure 8. Posterior model weights for (a) full observation time series and b) reduced observation time series as obtained from the different BME evaluation methods for models mHM1L (dark blue) and mHM2L (light blue). Green vertical line indicates reference solution generated by MC integration.

research. In summary, overly decisive weights are not an artifact of BME evaluation via ICs. They are a characteristic of BMA theory (and hence also of BME evaluation via numerical methods). Future investigations on the adequate choice of likelihood functions should therefore consider a suitable numerical method as a reference for the true characteristics of the BMA framework.

We hypothesize that, if the difference in model performance was not as dominant, model ranking would not be as obvious and not as accurately reproduced by approximation methods. To test this hypothesis, we shorten the observation time series in the spirit of a thought experiment and only consider the first year for calibration. By virtue of reducing the number of data points, the difference in goodness of fit will not be as pronounced because the multiplication of likelihoods for each data point of the time series will have a much less drastic effect. We do not advocate to use such a short time series for proper hydrological model calibration, but proceed with it in section 5.3 to illustrate the impact of the number of observations. Note that we do not change the analysis framework here (neither the definition of the likelihood function nor the formulation of how to calculate model weights from ICs) as suggested by the authors listed above, but actually change the amount of information that goes into the BMA analysis. Using the reduced time series for model selection can also be seen as representative for more difficult model selection tasks in other disciplines that have to cope with a limited number of measurements.

5.3. Results for Reduced Observation Time Series

When using the reduced observation time series of 365 daily discharge measurements for calibration, BME takes values of -1134 for the single-layer model (mHM1L) and -1132 for the double-layer model (mHM2L) in the reference solution. As intended, the difference in goodness of fit and therefore in BME values is now much less pronounced. Again, the numerical methods and the KIC achieve relatively good results. The AIC, AICc (considered here since $N_s/N_p < 40$), and BIC again drastically underestimate the reference values. Nested sampling consistently yields the most accurate BME value. Approximation results are summarized in Table 3.

This setup now leads to a less decisive weighting of 16.5% (mHM1L) versus 83.5% (mHM2L) in the reference solution. Apparently, the difference in performance over the reduced time series is now small enough such that model mHM1L should not be discarded, despite its slightly larger parameter space. Nevertheless, mHM2L still obtains a significantly larger posterior probability of being the better model due to its more

suitable representation of water retention capacity and water movement in the unsaturated zone. This time, however, not only the BIC yields model weights erroneously in clear favor of mHM1L, but also the AIC and AICc (even though they are derived for approximating BMA weights). Despite its BME approximation quality being inferior to nested sampling, posterior sampling surprisingly yields weights closest to the reference solution. Nested sampling and importance sampling weights still convey a similar message, but already deviate significantly from the reference weighting. Here the KIC yields indecisive weights of 51.9% versus 48.1% and therefore does not qualify as a reliable approximation method. Resulting model weights are visualized in Figure 8b.

5.4. Conclusions From Real-World Test Case

From this real-world test case, we have found that, still, the KIC performs best out of the ICs considered here, but it produces a much larger approximation error than any of the numerical methods. All of the methods used in this study, except for the BIC, are able to reproduce the reference model weighting when using a large calibration data set. The BIC yields exactly the opposite model ranking and therefore contradicts the true Bayesian model ranking, even in such a good-natured model selection problem where a large data set is available. The fact that the BIC did not perform particularly badly (e.g., definitely not worse than the AIC) in the synthetic setup (section 4) is alarming, because this means that there is not necessarily a correlation between the ICs' performance under linear conditions with their performance under nonlinear (but still good-natured) conditions. When using a significantly shorter time series, the quality of BME approximation deteriorates for the AIC and improves in case of the BIC, while the approximation quality of the KIC slightly decreases. This was expected from our investigation of the influence of the data set size in the synthetic test case (see section 4.3). Independently of the actual BME approximation quality, all ICs yield inaccurate posterior model weights. We therefore have to face the fact that the ICs considered here produce a rather arbitrary model ranking result in nonlinear real-world applications, and their accuracy cannot be predicted in a general manner. Only the numerical methods are able to reproduce the true model ranking reliably and sufficiently well.

6. Summary and Conclusions

In this study, we have compared nine methods to approximate Bayesian model evidence (BME), which is required to perform Bayesian model averaging (BMA) or Bayesian model selection. Since analytical solutions only exist under strongly limiting assumptions, we have investigated the usefulness of four numerical methods (simple MC integration, MC integration with importance sampling, MC integration with posterior sampling, and nested sampling) which do not rely on any assumptions, but suffer from high computational effort and potential inefficiency in high parameter dimensions. We have further considered four different mathematical approximations which are known as information criteria (AIC, AICc, BIC, and KIC evaluated at the MLE) and are frequently used in the context of BMA, but in previous studies yielded contradicting results with regard to model ranking. To be most consistent with approximation theory, we have proposed to evaluate the KIC at the MAP, instead, and also included this variant in our intercomparison. The nine BME evaluation methods analyzed in this study are summarized in Table 4.

6.1. Summary of Results

We have systematically compared these nine approximation techniques with regard to their theoretical derivations, common features, and differences in underlying assumptions. From this extensive analysis, we conclude that out of the ICs, the KIC evaluated at the MAP (KIC@MAP) is the most consistent one with BMA theory, but also the most expensive one to evaluate. It yields the true solution if the posterior parameter distribution is Gaussian (e.g., if the data set is large). The other ICs considered here are simplified versions of the KIC@MAP (KIC@MLE and BIC) or derived in a non-BMA context (AIC and AICc). Since the ICs' assumptions to calculate BME are too strong for nonlinear models or lack the correct theoretical foundation altogether, the contradicting reports on their performance at various accounts in the literature were to be expected. The numerical methods considered here are not limited by any assumptions. Simple MC integration is bias-free, but computationally very expensive. The other numerical methods are potentially more efficient, but prone to show a bias in their BME estimate. The most important assumptions and limitations of the nine BME evaluation methods analyzed in this study are summarized in Table 4.

The main contribution of this study is a first-time benchmarking of the different methods on a simplistic synthetic example where an exact analytical solution exists. The benchmarking has shown that all ICs (except for the KIC@MAP) potentially yield unacceptably large errors in BME approximation, with their performance depending on the actual calibration data set. Therefore, it cannot be recommended to use either of these criteria for a reliable, accurate approximation of BME. Especially, the AIC(c) and BIC cannot distinguish models which only differ in their prior definition of the parameter space. This is a major concern, since the specification of prior information is a fundamental part of Bayesian inference and Bayesian model ranking. It is misleading to think that the AIC(c) or BIC would perform acceptably if prior information is vague or not available; we have demonstrated that actually the opposite is the case. Also, these criteria are incapable of capturing the true dimensionality of a model. It remains an open question for future investigation, whether model dimensionality could be more adequately "encoded" in these ICs to improve their so far unacceptable performance in BME approximation. While the KIC@MAP represents a perfect choice if its assumptions are fulfilled, its performance deteriorates significantly with increasing nonlinearity in the considered models as illustrated in the second part of our synthetic example. The accuracy of BME approximation by the numerical methods considered here is only limited by computational effort (except for MC integration with posterior sampling, which is expected to overestimate BME due to its theoretical formulation). Simple MC integration and MC integration with importance sampling showed the highest accuracy and lowest uncertainty for a given computational effort, which agrees with the expectations based on the theoretical background of the respective methods.

We have continued our analysis with a real-world application to hydrological model selection. We have compared two conceptually slightly different versions of the distributed mesoscale hydrologic model (mHM) [*Samaniego et al.*, 2010], which are applied to the Fils river catchment of the Upper Neckar basin in Southwest Germany. Here the nonlinearity in model equations represents a typical case where the assumptions of the analytical solution or the Laplace approximation are not fulfilled. Based on the findings from our theoretical comparison and the benchmarking, we have therefore used simple MC integration as reference solution. In this realistic setup, the KIC again performs best out of the ICs considered in this study, but it produces a much larger approximation error than any of the numerical methods. Using a long time series of daily discharge measurements between 1980 and 1988 for calibration, model choice turns out to be very clear, such that the conceptually superior model obtains a posterior model weight of 100%. In this case, all of the methods considered in this study, except for the BIC, reproduce this clear model weighting, despite their difficulties in approximating the actual value of BME. Note that the BIC yields exactly the opposite model ranking and therefore contradicts the true Bayesian model ranking, even in this good-natured model selection problem where a large data set is available. The BIC's much worse performance in the real-world application could not be foreseen from its performance in the synthetic setup. This fact is revealing, because it reminds us again that the approximation quality of the ICs is rather arbitrary and application-dependent, which has already become evident in the synthetic test case.

We have pointed out that BMA theory supports the finding from our studies and previous work that, if only enough data points are available for calibration, one model will be the clear winner, even if the competing model is also performing very well as in our application. This might seem counter-intuitive, but is not an artifact of approximating BME via ICs as previously suspected by other authors. To test our hypothesis that also the other ICs discussed here would not be able to yield a consistent ranking if the calibration time series were shorter (as would often be the case for model selection in disciplines other than hydrology), we have repeated the analysis using only the first year of measurements for calibration. We do not advocate to use such a short data set for a robust calibration of mHM, but provide this analysis to make the reader aware of the important role of the data set length. Independently of the actual BME approximation quality, all ICs yield inaccurate posterior model weights in this case, while the numerical methods are able to reproduce the true model ranking sufficiently well.

6.2. Implications for Robust Model Selection

Both test case applications have revealed that using ICs to approximate BME potentially can, but not necessarily will, produce acceptable results for both the absolute BME value and the resulting model ranking. The Bayesian trade-off between model performance and model complexity is not represented adequately by the ICs, with the potential exception of the KIC@MAP. However, it cannot be decided in advance if a data set is large enough for the KIC@MAP to perform well for a given application and model set. This is why we

advocate to perform a comparison of the KIC@MAP with at least one of the numerical methods presented here, in order to assess the degree of agreement between the methods. If the discrepancy is large, one should continue with numerical methods if computationally feasible. If computation time is limited, nested sampling could be an efficient alternative to full MC integration as indicated by our test case results. If numerical evaluation is not an option at all because large model run times prohibit such an approach, we still do not recommend to compare several ICs amongst each other as frequently suggested in the literature, since this procedure cannot provide any conclusive insight. Instead, we suggest to solely use the KIC, evaluated at the MAP. Note that this still involves nonnegligible computational effort, since the MAP and the posterior covariance need to be determined; this could however be done by more efficient numerical optimization schemes. Nonetheless, finding a reliable alternative to numerical BME evaluation is still an open research question.

6.3. Conclusions

In conclusion, the findings from our theoretical intercomparison, the benchmarking results from our synthetic study as well as the insights from the application to a real-world hydrological model selection problem demonstrate that

1. for real-world applications, BME typically needs to be evaluated numerically or approximated by ICs because no analytical solution exists;
2. out of the ICs, the KIC evaluated at the MAP is the most consistent one, but might still be heavily biased when applied to nonlinear models;
3. the choice of evaluation method for BME substantially influences the deviation from the true BME value, the outcome of posterior model weights and model ranking as such;
4. for reliable model selection, there is still no reliable alternative to bias-free numerical methods.

Acknowledgments

The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the International Research Training Group "Integrated Hydrosystem Modelling" (IRTG 1829) at the University of Tübingen and the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart. This work was also supported by a grant from the Ministry of Science, Research and Arts of Baden-Württemberg (AZ Zu 33–721.3-2) and the Helmholtz Centre for Environmental Research, Leipzig (UFZ). The authors also thank R. Kumar for setting up the mHM for the Neckar basin. The authors further acknowledge the collaboration of the German Weather Service DWD for providing the observational data set employed in this study, which can be obtained via the WebWerdis data portal (www.dwd.de/webwerdis). The discharge data were obtained from LUBW Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg (www.lubw.baden-wuerttemberg.de/servlet/is/35855/). The data are "open to non-commercial research or training institutes, federal or state authorities and national meteorological services, which all fulfill the requirements of receiving the data free of charge."

References

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

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki, pp. 367–281, Springer, N. Y.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Automatic Control*, *19*(6), 716–723, doi:10.1109/TAC.1974.1100705.

Angluin, D., and C. H. Smith (1983), Inductive inference: Theory and methods, *ACM Comput. Surv.*, *15*(3), 237–269.

Bergström, S., B. Carlsson, G. Grahn, and B. Johansson (1997), A more consistent approach to catchment response in the HBV model, *Vannet i Norden*, *4*, 1–7.

Beven, K., and A. Binley (1992), The future of distributed models-–Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*(3), 279–298, doi:10.1002/hyp.3360060305.

Box, G., and G. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Mass.

Box, G. E. P., and D. R. Cox (1964), An analysis of transformations, *J. R. Stat. Soc., Ser. B*, *26*(2), 211–252.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997), Model selection: An integral part of inference, *Biometrics*, *53*(2), 603–618, doi:10.2307/2533961.

Burnham, K. P., and D. R. Anderson (2003), *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach*, vol. XXVI, 2nd ed., [corr. print.] ed., 488 pp., Springer, N. Y.

Burnham, K. P., and D. R. Anderson (2004), Multimodel inference-–Understanding AIC and BIC in model selection, *Sociol. Methods Res.*, *33*(2), 261–304, doi:10.1177/0049124104268644.

Chamberlin, T. C. (1890), The method of multiple working hypotheses, *Science*, *15*(366), 92–96.

Claeskens, G., and N. L. Hjort (2008), *Model Selection and Model Averaging, Cambridge Ser. Stat. Probab. Math.*, vol. 330, Cambridge Univ. Press, Cambridge, U. K.

Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.

De Bruijn, N. (1961), *Asymptotic Methods in Analysis, Bibliotheca Mathematica*, vol. 4, Dover, Amsterdam, Netherlands.

DeGroot, M. H. (1970), *Optimal Statistical Decisions*, McGraw-Hill, N. Y.

Del Giudice, D., M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann (2013), Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, *Hydrol. Earth Syst. Sci.*, *17*(10), 4209–4225, doi:10.5194/hess-17-4209-2013.

Efron, B. (1979), Bootstrap methods: Another look at the jackknife, *Ann. Stat.*, *7*(1), 1–26.

Elsheikh, A. H., M. F. Wheeler, and I. Hoteit (2013), Nested sampling algorithm for subsurface flow model selection, uncertainty quantification, and nonlinear calibration, *Water Resour. Res.*, *49*, 8383–8399, doi:10.1002/2012WR013406.

Evans, M., and T. Swartz (1995), Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems, *Stat. Sci.*, *10*(3), 254–272.

Evin, G., M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera (2014), Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resour. Res.*, *50*, 2350–2375, doi:10.1002/2013WR014185.

Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, vol. 1, 3rd ed., John Wiley, N. Y.
Fisher, R. A. (1922), On the mathematical foundations of theoretical statistics, *Philos. Trans. R. Soc. London A*, *222*, 309.
Foglia, L., S. W. Mehl, M. C. Hill, and P. Burlando (2013), Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland, *Water Resour. Res.*, *49*, 260–282, doi:10.1029/2011WR011779.
Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*(4), 457–472.
Geman, S., E. Bienenstock, and R. Doursat (1992), Neural networks and the bias/variance dilemma, *Neural Comput.*, *4*(1), 1–58, doi:10.1162/neco.1992.4.1.1.
Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resour. Res.*, *48*, W08301, doi:10.1029/2011WR011044.
Hammersley, J. M. (1960), Monte Carlo methods for solving multivariable problems, *Ann. N. Y. Acad. Sci.*, *86*(3), 844–874, doi:10.1111/j.1749-6632.1960.tb42846.x.
Hammersley, J. M., D. C. Handscomb, and G. Weiss (1965), Monte Carlo methods, *Phys. Today*, *18*, 55.
Härdle, W. (1991), *Smoothing Techniques: With Implementation in S, Springer Ser. Stat.*, Springer, N. Y.
Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, *57*(1), 97–109, doi:10.2307/2334940.
Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–401.
Hurvich, C. M., and C. L. Tsai (1989), Regression and time-series model selection in small samples, *Biometrika*, *76*(2), 297–307, doi:10.2307/2336663.
Jensen, J. L. W. V. (1906), Sur les fonctions convexes et les ingalits entre les valeurs moyennes, *Acta Math.*, *30*(1), 175–193, doi:10.1007/bf02418571.
Kashyap, R. L. (1982), Optimal choice of AR and MA parts in autoregressive moving average models, *IEEE Trans. Pattern Anal. Mach. Intell.*, *PAMI-4*(2), 99–104.
Kass, R. E., and A. E. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, *90*(430), 773–795, doi:10.2307/2291091.
Koehler, A. B., and E. S. Murphree (1988), A comparison of the Akaike and Schwarz criteria for selecting model order, *J. R. Stat. Soc. Ser. C*, *37*(2), 187–195, doi:10.2307/2347338.
Kuha, J. (2004), AIC and BIC: Comparisons of assumptions and performance, *Sociol. Methods Res.*, *33*(2), 188–229, doi:10.1177/0049124103262065.
Kumar, R., L. Samaniego, and S. Attinger (2010), The effects of spatial discretization and model parameterization on the prediction of extreme runoff characteristics, *J. Hydrol.*, *392*(1-2), 54–69, doi:10.1016/j.jhydrol.2010.07.047.
Li, X., and F. T.-C. Tsai (2009), Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod, *Water Resour. Res.*, *45*, W09403, doi:10.1029/2008WR007488.
Liang, X., E. F. Wood, and D. P. Lettenmaier (1996), Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification, *Global Planet. Change*, *13*(1-4), 195–206, doi:10.1016/0921-8181(95)00046-1.
Link, W. A., and R. J. Barker (2006), Model weights and the foundations of multimodel inference, *Ecology*, *87*(10), 2626–2635.
Liu, J. S. (2008), *Monte Carlo Strategies in Scientific Computing*, Springer, N. Y.
Lu, D., M. Ye, and S. P. Neuman (2011), Dependence of Bayesian model selection criteria and Fisher information matrix on sample size, *Math. Geosci.*, *43*(8), 971–993, doi:10.1007/s11004-011-9359-0.
Lu, D., M. Ye, P. D. Meyer, G. P. Curtis, X. Q. Shi, X. F. Niu, and S. B. Yabusaki (2013), Effects of error covariance structure on estimation of model averaging weights and predictive performance, *Water Resour. Res.*, *49*, 6029–6047, doi:10.1002/wrcr.20441.
MacKay, D. J. C. (1992), Bayesian interpolation, *Neural Comput.*, *4*, 415–447, doi:10.1162/neco.1992.4.3.415.
Meyer, P. D., M. Ye, M. L. Rockhold, S. P. Neuman, and K. J. Cantrell (2007), Combined estimation of hydrogeologic conceptual model, parameter, and scenario uncertainty with application to uranium transport at the Hanford Site 300 Area, *Tech. Rep. PNNL-16396*, Pac. Northwest Natl. Lab., Richland, Wash.
Morales-Casique, E., S. P. Neuman, and V. V. Vesselinov (2010), Maximum likelihood Bayesian averaging of airflow models in unsaturated fractured tuff using Occam and variance windows, *Stochastic Environ. Res. Risk Assess.*, *24*(6), 863–880, doi:10.1007/s00477-010-0383-2.
Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, and D. A. Stainforth (2004), Quantification of modelling uncertainties in a large ensemble of climate change simulations, *Nature*, *430*(7001), 768–772, doi:10.1038/nature02771.
Najafi, M. R., H. Moradkhani, and I. W. Jung (2011), Assessing the uncertainties of hydrologic model selection in climate change impact studies, *Hydrol. Processes*, *25*(18), 2814–2826, doi:10.1002/hyp.8043.
Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, *10*, 282–290.
Neuman, S. P. (2002), Accounting for conceptual model uncertainty via maximum likelihood Bayesian model averaging, in *Proceedings of ModelCARE2002*, vol. 277, pp. 303-313, IAHS Publ., Prague, Czech Republic.
Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environ. Res. Risk Assess.*, *17*(5), 291–305, doi:10.1007/s00477-003-0151-7.
Neuman, S. P., L. Xue, M. Ye, and D. Lu (2012), Bayesian analysis of data-worth considering model and parameter uncertainties, *Adv. Water Resour.*, *36*, 75–85, doi:10.1016/j.advwatres.2011.02.007.
Newton, M. A., and A. E. Raftery (1994), Approximate Bayesian inference with the weighted likelihood bootstrap, *J. R. Stat. Soc. Ser. B*, *56*(1), 3–48.
Poeter, E., and D. Anderson (2005), Multimodel ranking and inference in ground water modeling, *Ground Water*, *43*(4), 597–605, doi:10.1111/j.1745-6584.2005.0061.x.
Popper, K. R. (1959), *The Logic of Scientific Discovery*, Basic Books, N. Y.
Raftery, A. E. (1995), Bayesian model selection in social research, *Sociol. Methodol.*, *25*, 111–163, doi:10.2307/271063.
Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*(5), 1155–1174, doi:10.1175/mwr2906.1.
Refsgaard, J. C., J. P. van der Sluijs, J. Brown, and P. van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Adv. Water Resour.*, *29*(11), 1586–1597, doi:10.1016/j.advwatres.2005.11.013.
Reichert, P., and J. Mieleitner (2009), Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters, *Water Resour. Res.*, *45*, W10402, doi:10.1029/2009WR007814.
Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.

Riva, M., M. Panzeri, A. Guadagnini, and S. P. Neuman (2011), Role of model selection criteria in geostatistical inverse estimation of statistical data- and model-parameters, *Water Resour. Res.*, *47*, W07502, doi:10.1029/2011WR010480.
Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resour. Res.*, *44*, W12418, doi:10.1029/2008WR006908.
Samaniego, L., R. Kumar, and S. Attinger (2010), Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, *Water Resour. Res.*, *46*, W05523, doi:10.1029/2008WR007327.
Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464, doi:10.1214/aos/1176344136.
Schweppe, F. C. (1973), *Uncertain Dynamic Systems*, Prentice Hall, Englewood Cliffs, N. J.
Singh, A., S. Mishra, and G. Ruskauff (2010), Model averaging techniques for quantifying conceptual model uncertainty, *Ground Water*, *48*(5), 701–715, doi:10.1111/j.1745-6584.2009.00642.x.
Skilling, J. (2006), Nested sampling for general Bayesian computation, *Bayesian Anal.*, *1*(4), 833–859.
Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter-estimation procedures for hydrologic rainfall-runoff models––Correlated and heteroscedastic error cases, *Water Resour. Res.*, *16*(2), 430–442, doi:10.1029/WR016i002p00430.
Sugiura, N. (1978), Further analysts of the data by Akaike' s information criterion and the finite corrections, *Commun. Stat.*, *7*(1), 13–26, doi:10.1080/03610927808827599.
Tierney, L., and J. B. Kadane (1986), Accurate approximations for posterior moments and marginal densities, *J. Am. Stat. Assoc.*, *81*(393), 82–86, doi:10.2307/2287970.
Troldborg, M., W. Nowak, N. Tuxen, P. L. Bjerg, R. Helmig, and P. J. Binning (2010), Uncertainty evaluation of mass discharge estimates from a contaminated site using a fully Bayesian framework, *Water Resour. Res.*, *46*, W12552, doi:10.1029/2010WR009227.
Tsai, F. T.-C., and X. Li (2010), Reply to comment by Ming Ye et al. on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window," *Water Resour. Res.*, *46*, W02802, doi:10.1029/2009WR008591.
Tsai, F. T.-C., and X. B. Li (2008), Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window, *Water Resour. Res.*, *44*, W09434, doi:10.1029/2007WR006576.
Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*, W01411, doi:10.1029/2005WR004838.
Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.
Wöhling, T., and J. A. Vrugt (2008), Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models, *Water Resour. Res.*, *44*, W12432, doi:10.1029/2008WR007154.
Wöhling, T., L. Samaniego, and R. Kumar (2013), Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment, *Environ. Earth Sci.*, *69*(2), 453–468, doi:10.1007/s12665-013-2306-2.
Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.
Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resour. Res.*, *44*, W03428, doi:10.1029/2008WR006803.
Ye, M., K. F. Pohlmann, J. B. Chapman, G. M. Pohll, and D. M. Reeves (2010a), A model-averaging method for assessing groundwater conceptual model uncertainty, *Ground Water*, *48*(5), 716–28, doi:10.1111/j.1745-6584.2009.00633.x.
Ye, M., D. Lu, S. P. Neuman, and P. D. Meyer (2010b), Comment on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window" by Frank T.-C. Tsai and Xiaobao Li, *Water Resour. Res.*, *46*, W02801, doi:10.1029/2009WR008501.

Journal of Hydrology 531 (2015) 96–110

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection

Anneli Schöniger [a,*], Walter A. Illman [b], Thomas Wöhling [c,d], Wolfgang Nowak [e]

[a] Center for Applied Geoscience, University of Tübingen, Tübingen, Germany
[b] Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, Ontario, Canada
[c] Water & Earth System Science (WESS) Competence Cluster, Institute for Geoscience, University of Tübingen, Tübingen, Germany
[d] Lincoln Environmental Research, Lincoln Agritech Limited, Hamilton, New Zealand
[e] Institute for Modelling Hydraulic and Environmental Systems (LS^3)/SimTech, University of Stuttgart, Stuttgart, Germany

ARTICLE INFO

Article history:
Available online 15 August 2015

Keywords:
Groundwater modeling
Hydraulic tomography
Geostatistics
Bayesian model averaging
Model selection
Model calibration

SUMMARY

Groundwater modelers face the challenge of how to assign representative parameter values to the studied aquifer. Several approaches are available to parameterize spatial heterogeneity in aquifer parameters. They differ in their conceptualization and complexity, ranging from homogeneous models to heterogeneous random fields. While it is common practice to invest more effort into data collection for models with a finer resolution of heterogeneities, there is a lack of advice which amount of data is required to justify a certain level of model complexity. In this study, we propose to use concepts related to Bayesian model selection to identify this balance. We demonstrate our approach on the characterization of a heterogeneous aquifer via hydraulic tomography in a sandbox experiment (Illman et al., 2010). We consider four increasingly complex parameterizations of hydraulic conductivity: (1) Effective homogeneous medium, (2) geology-based zonation, (3) interpolation by pilot points, and (4) geostatistical random fields. First, we investigate the shift in justified complexity with increasing amount of available data by constructing a *model confusion matrix*. This matrix indicates the maximum level of complexity that can be justified given a specific experimental setup. Second, we determine which parameterization is most adequate given the observed drawdown data. Third, we test how the different parameterizations perform in a validation setup. The results of our test case indicate that aquifer characterization via hydraulic tomography does not necessarily require (or justify) a geostatistical description. Instead, a zonation-based model might be a more robust choice, but only if the zonation is geologically adequate.

1. Introduction

Groundwater models are built for various types of investigations, both in science and in practice. They can serve as a basis for hypothesis testing, risk assessment, and management of resources. To provide reliable predictions for these objectives, models must be calibrated sufficiently well. However, in light of limited budgets, modelers have to cope with small calibration data sets. For physically-based models that consider the fundamentally important processes, the calibration procedure aims at finding appropriate parameterizations and then constraining the plausible parameter ranges. In groundwater modeling, the most effort is typically spent on characterizing the heterogeneity of the subsurface parameters hydraulic conductivity and specific storage. Under steady-state assumptions, only the spatial distribution of hydraulic conductivity influences the flow conditions.

Several approaches are available to characterize the heterogeneity in hydraulic conductivity, which differ in effort and scale. Traditionally, a large number of hydraulic conductivity estimates is obtained from collecting core samples and performing permeameter tests (Sudicky, 1986; Sudicky et al., 2010), or from performing slug or pumping tests. The local-scale information obtained from such campaigns is then regionalized to larger scales by upscaling, zonation, interpolation, or geostatistical simulation. Alternatively, more detailed measurements can be obtained from geophysical investigations (e.g., Hubbard and Rubin, 2000) or hydraulic tomography (e.g., Gottlieb and Dietrich, 1995; Butler et al., 1999; Yeh and Liu, 2000; Straface et al., 2007; Li et al., 2007; Illman et al., 2010).

Hydraulic tomography has been developed to investigate the heterogeneity in aquifer properties in a fine spatial resolution. A number of pumping tests is performed sequentially in different

* Corresponding author.
E-mail address: anneli.schoeniger@uni-tuebingen.de (A. Schöniger).

http://dx.doi.org/10.1016/j.jhydrol.2015.07.047
0022-1694/

wells at various locations throughout the aquifer. Pumping induces a spatial distribution of drawdown, which is captured by observation wells throughout the domain. These drawdown data are then used to derive via numerical inversion the spatial distribution of hydraulic conductivity and related properties such as connectivity. Further, the uncertainty attached to the inferred parameters can be quantified. The spatial resolution of the derived parameter distribution depends on the horizontal well spacing and the vertical packer intervals (Yeh and Liu, 2000).

Several approaches exist for the analysis and interpretation of the data obtained from all these aquifer characterization methods, and for the representation of the observed spatial heterogeneity in groundwater models. In general, a groundwater model with a specific spatial structure of hydraulic conductivity must be assumed. These assumptions vary in their conceptualization and their complexity (e.g., the number of parameters involved). Please note that the definition of model complexity is not unique, ranging from pure parameter counting over factor analysis to concepts that take into account probability distributions of parameters, data-parameter sensitivity and predictive variance. In principle, any parameterization ranging from the simple homogeneous case with an effective conductivity value to a geostatistical random field could be used. For the inversion of hydraulic tomography data, geostatistics-based inverse modeling methods are most frequently applied, such as the quasi-linear geostatistical approach (QL) (Kitanidis, 1995) and the sequential successive linear estimator (SSLE) (e.g., Yeh and Liu, 2000).

Eventually, the adequacy of the inferred hydraulic conductivity field and the overall groundwater model will depend on both the aquifer characterization technique and the chosen parameterization. The more data are available for calibration, the more detailed heterogeneities can be resolved. While it is common practice to invest more effort into data collection for geostatistical models (e.g. in form of hydraulic tomography data) than for simpler, effective conductivity models (e.g. in the form of core samples, slug tests or single-hole tests), there is a lack of advice, which amount and information content of data is required to justify a certain level of model complexity. We therefore see a need for a method that balances calibration effort (meaning both the effort for data collection and the computational effort to perform the inversion with the model) with model complexity and, implicitly, with model predictive performance. Assuming that the calibration effort increases with data set size, we use the amount of available data as proxy for the calibration effort in the following.

The formal statistical approach of Bayesian model averaging (BMA) (Draper, 1995; Hoeting et al., 1999) qualifies as such a method. It objectively ranks a number of competing models based on their fit to available data. Starting from a prior belief about the plausibility of each considered model, BMA updates this belief with knowledge from observed data via Bayes' theorem, and yields posterior model probabilities that reflect the updated plausibility. These probabilities allow for a quantitative ranking of the competing models and provide a basis for model selection. If more than one model obtains a significant model probability, their predictions can be combined in a weighted average that uses the probabilities as model weights. Finally, the uncertainty caused by the inability to uniquely choose only one of the considered models can be quantified as between-model variance.

BMA has been used in various disciplines as a statistical tool for model averaging (e.g., Ajami and Gu, 2010; Najafi et al., 2011; Seifert et al., 2012), model selection (e.g., Raftery, 1995; Huelsenbeck et al., 2004), quantification of model choice uncertainty (e.g., Rojas et al., 2008; Singh et al., 2010; Troldborg et al., 2010; Ye et al., 2010), data worth analysis (e.g., Rojas et al., 2010; Neuman et al., 2012; Xue et al., 2014; Wöhling et al., 2015), and model component dissection (Tsai and Elshall, 2013; Elshall and Tsai, 2014). In groundwater modeling, it has been applied to choose between different parameterizations of aquifer heterogeneity, e.g. by Ye et al. (2004), Tsai and Li (2008), Rojas et al. (2008), Morales-Casique et al. (2010), Seifert et al. (2012), and Elsheikh et al. (2013), to name only a few selected examples. Refsgaard et al. (2012) provide a review of strategies, including BMA, to address geological uncertainty in groundwater flow and transport modeling.

In the context of groundwater model selection and calibration, finding a balance between performance and complexity is of great interest (e.g., Yeh and Yoon, 1981; Fienen et al., 2009; Elsheikh et al., 2013). BMA is ideally suited to guide this search, because it implicitly honors the principle of parsimony or "Occam's razor" (Jeffreys, 1939; Gull, 1988). The BMA ranking reflects an optimal tradeoff between goodness-of-fit and model complexity, with model complexity being encoded in the prior probability distributions of the model parameters. The prior uncertainty in parameters is propagated through the model to the predictions, which are then compared to the observed data. A wide predictive distribution will be penalized by BMA, whereas a precise and accurate predictive distribution will be favored.

Although this optimal tradeoff is a main result of BMA, BMA has not yet been used to find the data amount required to justify a given level of complexity. In a certain sense, this reverses the direction in which BMA is usually applied, i.e. to rank models of different complexity for a given data set. We intend to fill this gap by isolating the complexity component of the tradeoff from its performance counterpart. We achieve this in a synthetic setup for BMA, where the models are mutually tested against their own predictions, instead of against real data. We introduce the concept of a *model confusion matrix*, which expresses how likely it is to identify the respective true model given the current experimental setup. We refer to this analysis as *model justifiability analysis*, because it reveals whether any specific level of complexity can be justified by the available amount and type of data (independent of the actually measured values) through the eyes of BMA. The question of justifiability is hence detached from the observed data values and becomes a function of the calibration effort only. Note that the calibration effort does not depend on the information content in the data (the effort for data collection is the same, no matter if the data turn out to be informative or not). The sensitivity of the model parameters to the data, on the other hand, has an impact on the outcome of BMA results and on the justifiability analysis.

While the *justifiability* analysis is based on the experimental design but not the actually measured data values, the *adequacy* of a model with regard to a specific prediction goal is defined by the tradeoff between complexity and performance in predicting the actually observed data values. The observations serve simultaneously as training and testing data for the specified model purpose. Hence, model adequacy as opposed to justifiability is assessed by the standard BMA routine based on the observed data. We therefore propose to perform BMA in a two-step procedure, running the synthetic justifiability analysis for the experimental setup first and determining the adequacy of each model in light of the observed data values in a second step that consists of the conventional BMA method. The results of the first step will then help to decide whether (a) the identified most adequate model is really the best choice given the current set of models, or (b) whether the identified model is only optimal given the currently too limited amount and information content of the data. The latter could occur when the available data do not allow to identify a more complex model among the model set, although the more complex model would actually be closer to the observed response of the system.

Further, the justifiability analysis can uncover the reasons for two models obtaining almost the same weight in the conventional

BMA analysis. The model confusion matrix reveals whether two models are actually very similar in their predictions, while the conventional BMA analysis cannot distinguish this case from the case of two models that by chance achieve a similar overall goodness-of-fit.

As application study for our approach, we use hydraulic tomography data to characterize a synthetic heterogeneous sandbox aquifer. The experimental setup of the sandbox and the data collection are described in Illman et al. (2010). We consider four conceptual models of vastly different structure and complexity to parameterize the spatial heterogeneity of hydraulic conductivity: a very simple homogeneous model, a zonation-based model that mimics the observed layering in the sandbox, and two variants of a highly flexible geostatistical model. For later analysis, we also use a geologically uninformed zonation model.

The prediction target of the groundwater model, using either one of these parameterizations, are the changes in groundwater head induced by sequential pumping at various locations throughout the sandbox aquifer. It is not clear beforehand which level of detail is required in the representation of heterogeneities to achieve reliable predictions. Since we know the true layering quite well from the experimental setup of the sandbox, the zonated model with its structure derived from the visible packing pattern might be seen as the favorite model. However, due to boundary and mixing effects between different sand types as well as errors involved with packing and with discretization, this is merely a hypothesis that remains to be tested. It is further a priori unknown whether the much more flexible geostatistical model should not be seen as the favorite.

In a parallel study, Illman et al. (2015) investigated the performance of a geostatistical model and a geologically well-informed zonation model when inverting hydraulic tomography data obtained from the same sandbox. They found that while geostatistics-based inversion yielded the best performance in terms of both calibration and validation results, the much less complex zonated model performed almost as well. The question how to choose optimally between these models in the spirit of Occam's razor will be answered in this study. We will perform the BMA analysis with an increasing amount of hydraulic tomography data from one to six cross-hole pumping tests. The resulting model rankings will point toward the most plausible parameterization in the spirit of Occam's razor as a function of available data.

With our proposed two-step procedure, we can answer the following two questions: (1) Which amount of hydraulic tomography data would be hypothetically needed to justify the use of a geostatistical approach for parameterizing the groundwater model? Or, vice versa, up to which amount of data should we choose a less flexible model instead, because it can still be calibrated reasonably well? And (2), which of the parameterizations is the most adequate one given the actually measured data?

Finally, we determine the predictions of the individual models and the BMA-weighted average for four independent cross-hole pumping tests not used in the BMA analysis. We compare the predictive performance of all models in this validation step to find out whether, in this specific test case, BMA is able to identify a robust model and whether the weighted average outperforms the optimal model.

As main contributions of this study, we demonstrate the application of BMA to choose between groundwater models which differ in the complexity of hydraulic conductivity parameterization by orders of magnitude. We introduce a model confusion matrix which indicates the maximum level of complexity that is theoretically justified by a specific experimental setup. This analysis serves as a basis for interpreting the model ranking which emerges from the actually observed data. Our suggested add-on to the standard BMA routine lays a special focus on the key ingredient of BMA, the implicit tradeoff between performance and complexity. As a side product, the model confusion matrix helps to revise the subjective choice of prior model probabilities, because it reveals the degree of similarity between the alternative models.

We briefly summarize the statistical framework of BMA and present our proposed model justifiability analysis in Section 2. In Section 3, we outline the general procedure of groundwater model calibration via hydraulic tomography, and provide details on the experimental setup, the hydraulic conductivity parameterizations, and the numerical implementation. Section 4 demonstrates the application of the justifiability analysis in a synthetic setup, while the model ranking based on the actually observed data is presented in Section 5. We summarize the insights from this study in Section 6.

2. Model ranking methodology

2.1. Bayesian model averaging

The mathematical framework of the BMA analysis is outlined and discussed comprehensively in Hoeting et al. (1999). We briefly present the relevant equations here. Note that all probabilities and statistics are conditional on the chosen set of models.

Based on the predictive distributions $p(\boldsymbol{\varphi} \mid \mathbf{y}_o, M_k)$ for a predicted quantity $\boldsymbol{\varphi}$ obtained from N_m competing models M_k, their weighted average according to BMA is given by

$$p(\boldsymbol{\varphi} \mid \mathbf{y}_o) = \sum_{k=1}^{N_m} p(\boldsymbol{\varphi} \mid \mathbf{y}_o, M_k) P(M_k \mid \mathbf{y}_o) \tag{1}$$

with $p(\cdot \mid \mathbf{y}_o)$ representing a probability distribution conditional on the observed data $\mathbf{y}_o$. $P(M_k \mid \mathbf{y}_o)$ is the posterior probability of model M_k to be the best one (in light of the given data $\mathbf{y}_o$) in the set of considered models. These posterior probabilities are used as weights in the model averaging procedure.

The posterior mean of the prediction $\boldsymbol{\varphi}$ is determined as the weighted average of the mean predictions by the individual models:

$$E[\boldsymbol{\varphi} \mid \mathbf{y}_o] = \sum_{k=1}^{N_m} E[\boldsymbol{\varphi} \mid \mathbf{y}_o, M_k] P(M_k \mid \mathbf{y}_o). \tag{2}$$

The posterior variance of the model-averaged predictive distribution is given by

$$V[\boldsymbol{\varphi} \mid \mathbf{y}_o] = \sum_{k=1}^{N_m} V[\boldsymbol{\varphi} \mid \mathbf{y}_o, M_k] P(M_k \mid \mathbf{y}_o) + \sum_{k=1}^{N_m} (E[\boldsymbol{\varphi} \mid \mathbf{y}_o, M_k] - E[\boldsymbol{\varphi} \mid \mathbf{y}_o])^2 P(M_k \mid \mathbf{y}_o) \tag{3}$$

with the first term representing within-model variance due to parameter uncertainty and the second term representing between-model variance due to conceptual uncertainty (uncertainty in model choice).

The posterior model weights $P(M_k \mid \mathbf{y}_o)$ are derived from Bayes' theorem. To obtain these weights, the prior belief about each model's adequacy $P(M_k)$ is updated with the evidence of the observed data:

$$P(M_k \mid \mathbf{y}_o) = \frac{p(\mathbf{y}_o \mid M_k) P(M_k)}{\sum_{i=1}^{N_m} p(\mathbf{y}_o \mid M_i) P(M_i)}, \tag{4}$$

where $p(\mathbf{y}_o \mid M_k)$ represents Bayesian model evidence (BME). It quantifies the average likelihood of the observed data for model M_k, accounting for its prior parameter space $\mathcal{U}_k$:

$$p(\mathbf{y}_o \mid M_k) = \int_{\mathcal{U}_k} p(\mathbf{y}_o \mid M_k, \mathbf{u}_k) p(\mathbf{u}_k \mid M_k) d\mathbf{u}_k. \tag{5}$$

Here, $p(\mathbf{u}_k \mid M_k)$ denotes the prior distribution of the parameters $\mathbf{u}_k$ in model M_k, while $p(\mathbf{y}_o \mid M_k, \mathbf{u}_k)$ is the likelihood of the observed data corresponding to the parameter set $\mathbf{u}_k$ of model M_k.

The integral in Eq. (5) can be either evaluated numerically (Kass and Raftery, 1995) at high computational costs, or approximated mathematically by so-called information criteria. The Kashyap information criterion (KIC) (Neuman, 2003), the Bayesian information criterion (BIC) (Schwarz, 1978; Raftery, 1995), and the Akaike information criterion (AIC) (Akaike, 1973) are the most commonly used ones within the BMA framework, although the AIC does not originate from the Bayesian context (e.g., Burnham and Anderson, 2004). While these criteria are favorable due to their computational efficiency, they have been shown to yield inaccurate and contradicting model ranking results in numerous studies (see e.g., Ye et al., 2008; Tsai and Li, 2008; Ye et al., 2010; Singh et al., 2010; Morales-Casique et al., 2010; Foglia et al., 2013). In a benchmarking exercise (Schöniger et al., 2014), these information criteria have been tested against computationally expensive numerical reference solutions for BME in both synthetic and real-world cases. The benchmarking clearly demonstrated that information criteria are poor approximations of BME in almost all cases and yield misleading model ranking results, and that brute-force numerical techniques for evaluating BME should be used whenever possible. Thus, we perform the BMA analysis in a Monte Carlo framework and evaluate BME via brute-force Monte Carlo integration. This method was shown to yield the most accurate results in the intercomparison study of Schöniger et al. (2014).

2.2. Model justifiability analysis

In this study, we focus on the properties of the BMA tradeoff between performance and complexity. In the limit of an infinite data set size, there is no need for such a tradeoff because the true model is perfectly known. Then, BMA will identify the true model (if it is part of the model set) with a weight of 100%, regardless of its complexity. We will call the successful identification of the true model through BMA *model self-identification*. With a limited amount of data, however, the optimal tradeoff has to be found, because an overly flexible model will fail in producing reliable prognoses (mostly due to over-fitting problems), while a too simple model will make overly confident, biased predictions.

Self-identification is achieved faster for low-complexity true models, because high complexity can only be justified by large data sets. We will prove that BMA is consistent with this reasoning by showing that the maximum level of complexity that can still be self-identified is shifting upwards with increasing data set size and downwards with decreasing data set size. Note that, in contrast to other methods that choose an optimal model based on a non-Bayesian tradeoff between performance and complexity, simple true models will still be recognized by BMA through a large calibration data set, while other methods such as the AIC tend to overestimate the true complexity with increasing data set size (Burnham and Anderson, 2004).

We propose to test how much data are required for model self-identification in a controlled, synthetic setup. We let the competing models take turns in generating synthetic data sets, based on random parameter realizations per model. Then, we perform the standard BMA analysis for each of the synthetic data sets, and average the obtained model weights over all data sets that were generated by a specific model. The averaged weights can be summarized in what we call a *model confusion matrix*. We borrow the term "confusion matrix" from the field of machine learning. A confusion matrix, or "contingency table", refers to a specific table layout that easily visualizes the score of a classifier algorithm by dividing into correctly classified objects and erroneously ("confused") classified objects. Such a matrix is quadratic and, in the case of a model confusion matrix, has the size $N_m \times N_m$.

Fig. 1 shows a schematic drawing of a model confusion matrix in its optimum state. The columns correspond to the respective defined-to-be-true models, i.e. the data-generating models. The

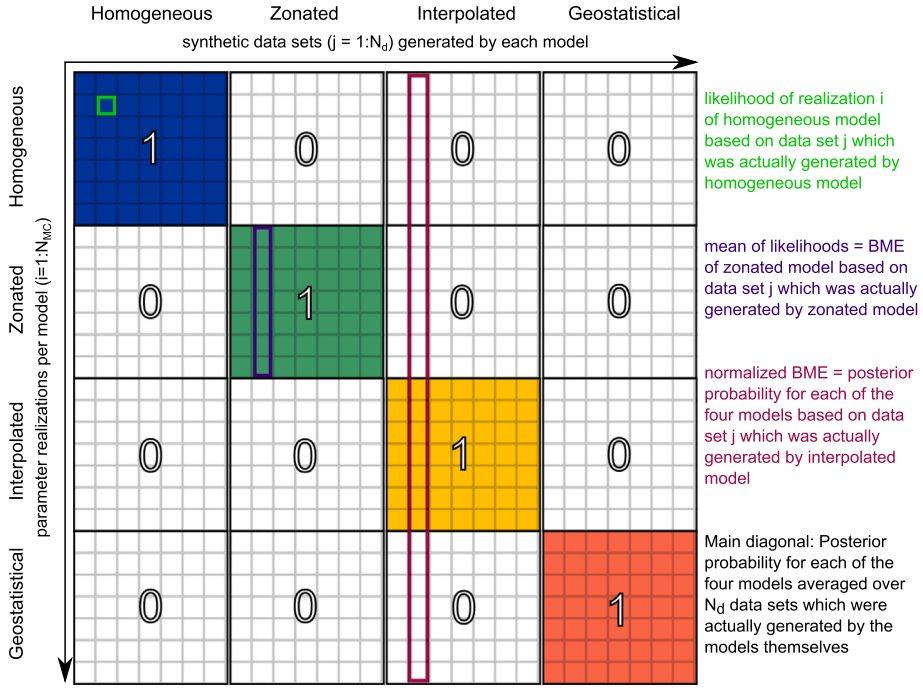


Fig. 1. Schematic illustration of how the model confusion matrix is constructed in a Monte Carlo framework for BMA. Each model takes its turn to generate N_d random synthetic data sets. For each of these data sets, model weights are determined based on the BME values achieved by the competing models. Model weights are averaged over the data sets generated by each model. The average weights for the four models are placed into the respective column that corresponds to the data-generating model. In the best case scenario, each model can be perfectly identified based on its own data sets, i.e. the main diagonal consists of mean model weights equal to one, and the off-diagonal entries are equal to zero.

rows correspond to the models that shall be ranked through BMA. Each data-generating model is perfectly identified as the true one with a posterior probability of 100% (all main diagonal entries are equal to one), whereas all models that did not generate the data receive a model weight of 0% (all off-diagonal entries are zero). For actual applications, however, the matrix is expected to show a suboptimal state. Models that did not generate the data could erroneously receive a significant probability of being the true one (they could be "confused" with the true one), while the true models might receive a weight significantly less than one. Note that this is equivalent to statistical hypothesis testing, with two competing models formulated as null hypothesis and alternative hypothesis, respectively. The model confusion matrix corresponds to the Bayesian probability matrix of correct outcomes versus type I and type II errors. However, in the model comparison case, typically there is no clear null hypothesis, such that none of the errors is perceived more critical than the other.

The model confusion matrix yields the maximum level of complexity that could potentially be self-identified and hence justified from the current experimental setup. Justifiability is achieved, if the data-generating model receives a higher weight than the competing candidate models. The absolute value of the weight relates to the degree of justifiability, i.e. a weight of one corresponds to perfect justifiability, and a weight of slightly more than $1/N_m$ corresponds to highly uncertain justifiability. The maximum level of complexity that can be self-identified corresponds to the complexity of the most complex model in the set that can still be justified.

If two or more models cannot be distinguished (i.e., the true model cannot be identified), this can have two reasons. First, the models might actually be very similar in their predictions. This means that, with the given experimental setup, they can hardly be discriminated. In that case, a modeler could decide to assign "diluted" prior weights in order to account for the fact that these models (seem to) belong to the same group of models (George, 2010). Second, the amount of data is too low to identify a complex true model because of the principle of parsimony. If the models cannot be sufficiently distinguished under the given conditions, other experimental designs (i.e., including more data and/or other data types) could be used to test whether the models at hand are actually not distinguishable or whether the tradeoff with parsimony under a limited amount of data was the reason for the lack in justifiability.

The model confusion matrix further represents the potential of each model in the model competition. A model that can be perfectly self-identified with the given experimental setup is expected to also obtain a weight of close to 100% in the model ranking based on actually observed data, if it represents the true system sufficiently well. Any significant decrease in the model weight when moving from the synthetic data analysis to the real data analysis indicates a significant mismatch between the model predictions and the data.

3. Groundwater model calibration via hydraulic tomography

We will base our analysis on experimental data from Illman et al. (2010). They used steady-state cross-hole pumping tests to perform hydraulic tomography of a sandbox aquifer. The experimental setup is briefly described in Section 3.1. For the inversion of the hydraulic tomography data, a groundwater model must be assumed. The four competing parameterizations of hydraulic conductivity we consider in the groundwater model are presented in Section 3.2. These four parameterizations constitute the set of models to be ranked by the BMA analysis in Section 5. Note that we use the terms "parameterization" and "hydraulic conductivity model" interchangeably in this study. Section 3.4 describes the inversion procedure, and Section 3.5 the numerical implementation of both the inversion and the BMA analysis.

3.1. Experimental setup

The construction of the synthetic aquifer in a vertical sandbox is described in detail in Illman et al. (2010), such that we will only summarize the main characteristics here. The sandbox dimensions are 193.0 cm length, 82.6 cm height and 10.2 cm width. A natural layering was achieved by filling the sandbox through a cyclic flux of sediment-laden water under varying conditions, using eleven different sand types. The final layering is visible from the frontal view photograph in Fig. 2. Hydraulic conductivity of the layers varies between less than 1×10^{-2} cm/s and more than 3×10^{-1} cm/s. This layering was constructed with natural sedimentation processes in mind, with the goal to test geostatistical inversion techniques on realistic structures.

At 48 ports, horizontal wells were installed which allow for monitoring with pressure transducers or for pumping. Since the wells penetrate the sandbox horizontally, the aquifer can be modeled in 2D, and the domain relevant for modeling is sized 160 cm $\times$ 78 cm. The hydraulic heads at the left- and right-hand boundaries are fixed by constant head reservoirs, and we treat the top boundary as a constant head boundary due to ponding water.

We use up to six cross-hole pumping tests for hydraulic tomography, i.e. for calibration and ranking of the groundwater models.

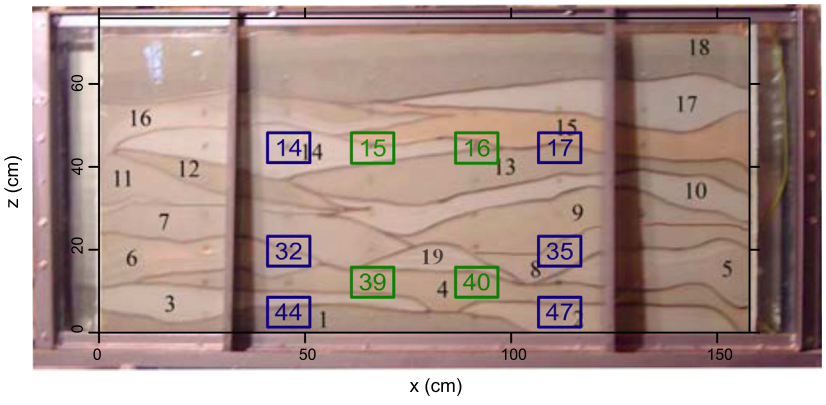


Fig. 2. Experimental setup of the sandbox aquifer (modified after Illman et al. (2010)): Black numbers indicate soil layers, blue and green numbers in squares indicate port numbers. Blue marked squares represent the ports used for hydraulic tomography (calibration), green marked squares represent ports used for independent cross-hole pumping tests (validation).

The corresponding ports are labeled in blue in Fig. 2. We further use pumping tests performed in four other ports (labeled in green in Fig. 2) to validate the conditioned model predictions.

We refer to Illman et al. (2010) for details on the available data besides the cross-hole pumping tests used for conditioning and model ranking here, such as hydraulic conductivity estimates from core samples or single-hole pumping tests.

3.2. Set of considered models

We consider four alternative parameterizations of hydraulic conductivity in the groundwater model: (1) A homogeneous, effective-value approach, (2) a zonation approach (with a variation in later analysis steps), (3) a geostatistical interpolation approach based on pilot points, and (4) a full geostatistical approach. These four parameterizations result from the following reasoning:

The simplest model is an effective-value approach, assuming an equivalent homogeneous medium. This model only uses one single parameter, which is the effective hydraulic conductivity value of the whole domain. Such a model is plausible if only very limited data is available to characterize the study area.

If knowledge about the site-specific geology is available, zones of constant hydraulic conductivity are typically derived. Here, we define the structure of the zonated model based on visual inspection of the sandbox layers. This of course represents a best-case scenario, since such detailed geological information is not available in field applications. We still include this approach to investigate the potential of geology-based zonation when calibrated with hydraulic tomography data. The zonated model consists of 19 zones (see numbered zones in Fig. 2), which corresponds to 19 parameters (one hydraulic conductivity value per zone).

The most complex parameterization is represented by a discretized fully geostatistical field, where each model cell (here: 12,480 grid cells, see Section 3.5) has a hydraulic conductivity value that can vary randomly according to a geostatistical model. Geostatistical inversion is the most commonly used approach when using hydraulic tomography data (e.g., Yeh and Liu, 2000; Li et al., 2005; Cardiff et al., 2009; Schöniger et al., 2012).

To soften the difference in complexity (assessed through parameter counting) between the zonated and the geostatistical model, we also consider a model variant which is generated by geostatistical interpolation between 120 pilot points (e.g., RamaRao et al., 1995; Vesselinov et al., 2001; Castagna and Bellin, 2009). The interpolated model only uses those cells as parameters which correspond to pilot point locations. This approach might be considered a reasonable compromise between the ability to fit the calibration data well and the potential over-parameterization by full geostatistics.

To visualize the structural differences between the different models, we show individual parameter realizations of each model in the left column of Fig. 4. These realizations represent conditional realizations when calibrating on the observations during the pumping test in port 44. Structural differences in the hydraulic conductivity fields are most obvious between the homogeneous model (Fig. 4a), the zonated model (Fig. 4b), and the two geostatistical approaches (Fig. 4c and d): While the zonated model shows characteristic edges along the zone boundaries, the geostatistics-based approaches show a smoother transition between small-scale areas of lower or higher hydraulic conductivity. The homogeneous approach obviously is maximally smooth. Differences between the two geostatistics-based approaches are less evident. The interpolated field is even more smooth than the fully geostatistical field, with more connected areas of high or low hydraulic conductivity. Resulting differences in drawdown predictions can be seen in the right column of Fig. 4. The drawdown patterns produced by the zonated model and the two geostatistics-based models look very similar, but differ from the pattern produced by the homogeneous model. We will analyze the performance of the four alternative parameterizations in detail in Section 5.

3.3. The problem of assessing model complexity and model justifiability

Intuitively, the four considered models vary greatly in complexity. However, there is no clear definition of complexity that would allow us to judge which one of those models is most appropriate given a specific amount of data. To illustrate the differences in complexity, we have chosen three different measures to quantify complexity: first, we simply count the number of adjustable parameters as, e.g., done by some of the information criteria (see Section 1); second, we use the results of a factor analysis based on the hydraulic conductivity fields produced by the models; and third, we determine the average standard deviation of drawdown predictions when calibrated on six pumping tests. The latter is most closely related to how BMA actually deciphers model complexity, because it is a measure for a model's flexibility in the prediction space as opposed to the parameter space. The three measures are shown in Fig. 3. The color scheme used here ranges from blue for lowest complexity to red for highest complexity and is maintained throughout the manuscript.

We note that, in factor analysis, the number of determined factors strongly depends on the used criterion. Here, we have applied the Kaiser criterion (Kaiser, 1960) which claims that a factor should explain a larger variation than an average single item. If we instead choose to keep all those factors which in sum explain 90% of the total parameter variance, the number of determined factors will change (see Fig. 3). Thus, the resulting number of factors can only provide a rough indication how many data points are needed when aiming at a calibration problem that should at least be determined (if not even overdetermined). Here, the factor analysis could not provide a clear hint whether the most complex geostatistical model is appropriate or over-parameterized given the at maximum 210 available data values.

While simply counting the number of parameters without accounting for correlation is obviously misleading, a factor analysis could be a useful pre-processing tool to make a first guess about the complexity and justifiability of a model with little effort. However, it cannot reveal anything about the merits of each model compared to each other and in light of the data. Considering all of parameter number, parameter variances and data-parameter

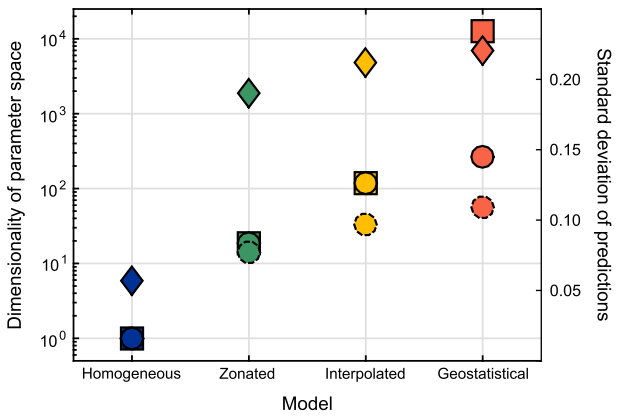


Fig. 3. Complexity of the four considered parameterizations, quantified as (a) the number of adjustable parameters (squares, left axis), (b) the number of factors with eigenvalues larger than one (circles, left axis), (c) the number of factors that explain 90% of the total variance (dashed circles, left axis), and (d) the average standard deviation of drawdown predictions (diamonds, right axis).

sensitivity to assess the justifiability of a model's complexity calls for a more elaborate approach such as BMA.

3.4. Inversion procedure

A review of inversion techniques for hydraulic tomography is given by Cardiff and Barrash (2011) and by Schöniger et al. (2012). Here, we solve the inversion problem by fully Bayesian updating of the prior parameter distribution with the evidence provided by the hydraulic tomography data. We obtain posterior parameter estimates with the bootstrap filter (Gordon et al., 1993). While computationally expensive, this algorithm does not require any assumptions (such as Gaussianity or linearity) on the involved probability distributions and the dependence between data and parameters.

3.5. Numerical implementation

For the numerical implementation of the inversion procedure, we discretize the model domain (the sandbox) into $160 \times 78 =$ 12,480 elements and 12,719 nodes. This discretization corresponds to a cell size of 1×1 cm.

To assign parameter values to each cell, we draw from prior parameter distributions. For all models, we assume that log-conductivity ($\ln K$) follows a Gaussian distribution. The prior mean is set to $\ln K = -2.56$ (corresponding to $K = 0.077$ cm/s), which is the average value of hydraulic conductivity estimates from core samples, single-hole pumping tests and one cross-hole pumping test not used in either calibration or validation here.

For the homogeneous medium approach, we draw random $\ln K$ values from a univariate Gaussian distribution centered about this prior mean, with a variance of $s^2_{\ln K} = 0.87$. This value corresponds to the variance observed in the core samples, which are deemed to represent local-scale variability best out of the available data types. Given a specific structure, the variance of the prior parameter distribution defines the flexibility of the model. By assigning a realistic value, we aim at making the BMA competition, which penalizes flexibility (complexity), as fair as possible.

For the zonated model, we draw random parameter values from the same distribution for each of the zones independently, i.e. we do not assume correlation across the layer boundaries. Thus, while we make use of the visible packing pattern in the sandbox to inform the geometry of the layers, the model does not benefit from any knowledge about the properties of each layer.

For both geostatistics-based parameterizations, we assume a stationary multi-Gaussian distribution for $\ln K$, with an exponential covariance model and a nugget of zero. The exponential variogram model is defined by the prior mean, the prior variance, and correlation lengths in horizontal and vertical directions, respectively. The prior mean is again set to $K = 0.077$ cm/s, to be consistent over all four parameterizations. The prior variance is increased to $s^2_{\ln K} = 2.0$, since we are trying to mimic the sharp edges between layers with a smooth spatial model. The correlation length in vertical direction is derived from visual inspection of the sandbox layers and set to 6 cm. The correlation length in horizontal direction is set to 60 cm. This value is determined from maximizing the likelihood of the hydraulic conductivity values obtained from core samples, given the predefined mean, variance, and correlation length in vertical direction. Note that we use other available data besides the cross-hole pumping tests for the definition of the prior parameter distributions, but not for direct conditioning of the spatial random fields, to maintain comparability with the other two parameterizations.

To set up the interpolated model, we define $6 \times 20 = 120$ pilot point locations on a regular grid. Their configuration is chosen such that each correlation length of the geostatistical model is sampled by about two pilot points. A random realization of this model is generated by drawing random values from the multivariate prior distribution for the pilot points. This approach still honors the spatial dependence between the pilot point values prescribed by the geostatistical model. Subsequently, the values in cells between the pilot points are interpolated based on kriging (e.g., Kitanidis, 1997). Recall that this type of interpolation is deterministic, which constitutes the difference to the fully geostatistical model. In the fully geostatistical case, each cell is allowed to vary randomly according to the geostatistical model. This is the reason why the interpolated model produces smoother spatial fields than the geostatistical model as observed in the left column of Fig. 4. The random fields are generated with the FFT-based random field generator also used in Nowak et al. (2008, 2010, 2012).

Between one and six pumping tests are included into the inversion (see ports marked blue in Fig. 2). For each pumping test, 48 observations of steady-state drawdown are available. Here, we exclude the observation directly at the pumping port. We further exclude observations from the top of the sandbox (observation ports 1 to 12, Illman et al. (2010)), where the signal-to-noise ratio was found to be very low in a pre-processing analysis not shown here. This leaves us with 35 observations per pumping test, yielding up to 210 observations used in the inversion.

We define the likelihood function $p(\mathbf{y}_o \mid M_k, \mathbf{u}_k)$ needed to perform the Bayesian update and to solve the Bayesian integral in Eq. (5) through a Gaussian distribution, centered about the observed data $\mathbf{y}_o$. We assume uncorrelated errors with a standard deviation of 1 cm. This assumption covers both measurement errors of up to 0.5 cm and unknown structural errors. In general, correlated errors and other distribution shapes should be used to define the likelihood function where adequate. Model-structural errors could also be considered with more complex statistical or stochastic descriptions, but such an analysis would be beyond the scope of the current study.

Based on random realizations generated from each of the four parameterizations as described above, we determine the BME value by brute-force Monte Carlo integration of Eq. (5); i.e., we average the likelihood values obtained from all parameter realizations.

We monitor the convergence of the BME estimate over increasing ensemble sizes to ensure a stable estimate as basis for the calculation of model weights. The final ensembles comprise 200,000 realizations for the homogeneous model, and 10 million realizations in the case of the zonated, the interpolated, and the geostatistical model.

We further determine the effective sample size (ESS) (Liu, 2008) for each model ensemble to detect any problems with degeneration of the bootstrap filter. The ESS measures the number of realizations which contribute significantly to the BME estimate and the posterior model predictions. ESS values vary between models and combinations of pumping tests used in the inversion. The lowest ESS results from calibration on all six pumping tests. ESS values then range between 11,671 for the homogeneous model and 171 for the geostatistical model. Together with the convergence analysis, these values confirm that we obtained sufficiently reliable posterior statistics.

4. Justifiability of the four parameterizations based on synthetic data

4.1. Construction of the model confusion matrix

To perform a model justifiability analysis, we first need to generate synthetic data sets with each one of the competing models. We therefore randomly draw $N_{d,k} = 1000$ parameter realizations

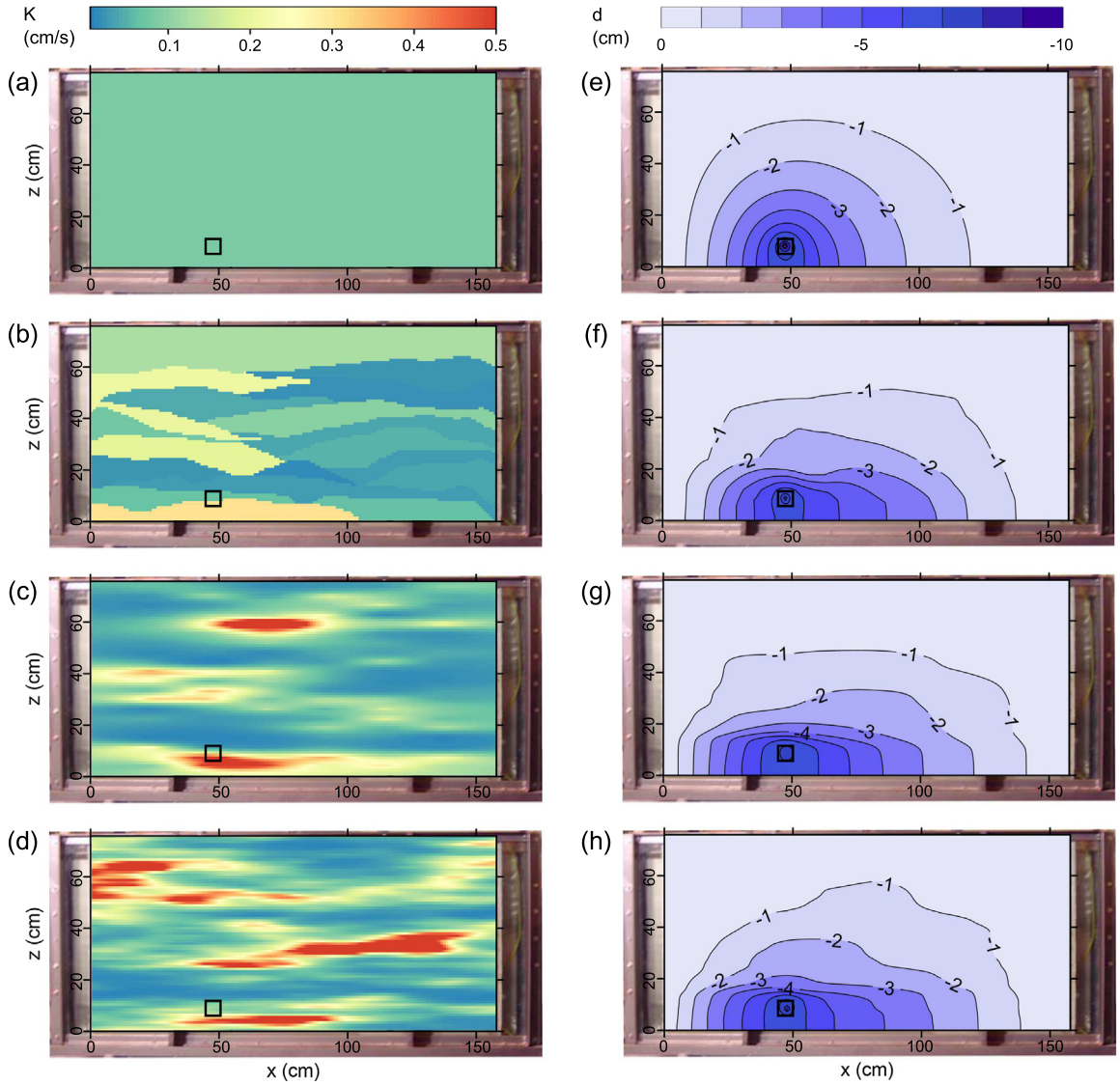


Fig. 4. Conditional realizations of hydraulic conductivity (left column) obtained from the four different parameterizations when conditioning on the observed drawdown during pumping test in port 44 (black square), and corresponding drawdown prediction (right column). (a, e) Homogeneous model; (b, f) zonated model; (c, g) interpolated model; (d, h) geostatistical model.

from the prior distribution $p(\mathbf{u}_k \mid M_k)$ for each of the four models and determine their corresponding drawdown predictions. We use the same experimental setup and configuration of observations for the synthetic truths as described for the real observations (see Section 3.1). Based on each of these synthetic data sets $\mathbf{y}_{o,k,j}$, $k = 1,\ldots,N_m$ and $j = 1,\ldots,N_{d,k}$, posterior model weights for all of the competing models are determined from Eq. (4). This sums up to a number of $N_d = \sum_{k=1}^{N_m} N_{d,k}$ BMA runs for a given experimental setup. The prior model weights are set to $P(M_k) = 1/N_m$, such that all models are equally likely before accounting for the evidence in the synthetic data sets. In a post-processing step, the posterior model weights are averaged over each data-generating model, i.e. all those weights are averaged that were obtained with the synthetic data sets generated by model M_l:

$$P(M_k \mid \mathbf{y}_{o,l}) = \frac{1}{N_{d,k}} \sum_{j=1}^{N_{d,k}} P(M_k \mid \mathbf{y}_{o,l,j}), \tag{6}$$

where $P(M_k \mid \mathbf{y}_{o,l})$ denotes the expected weight for model M_k under data from model M_l.

Fig. 1 schematically shows how the model confusion matrix is constructed and what its entries represent. The rows correspond to models M_k (i.e. homogeneous, zonated, interpolated, and geostatistical model, respectively) for which a posterior probability has been determined based on integration of likelihoods over N_{MC} prior parameter realizations (Monte Carlo integration of Eq. (5)). The columns correspond to models M_l (the same set as M_k) which generated the synthetic data. Each cell of this matrix shows the expected probability according to Bayes' theorem that, out of the considered set of models, model M_k is perceived as being most adequate to predict a data set which was actually generated from model M_l. Thus, the averaged model weights in each column sum up to one.

To investigate the influence of data set size on the outcome of model justifiability, we repeat the analysis for one, two, three, four, five, and six pumping tests included in the inversion. Considering all possible combinations of pumping test locations (e.g., 15 in the case of two pumping tests, 20 in the case of three pumping tests, and so on), this adds up to $N_v = 63$ data set variants. For each of the 63 data set variants, we obtain a model confusion matrix. Finally, we average all model confusion matrices that correspond to a specific data set size and assess their (dis-) similarity with

the optimal result as given by the identity matrix (see explanations in Section 2.2).

In total, $N_v \cdot N_d = 252,000$ BMA runs are performed. Compared to the effort required for the inversion (i.e., the effort to generate the prior ensemble of parameter realizations to be used for bootstrap filtering, see Section 3.4), additional BMA runs are relatively cheap, because they only need evaluations of likelihoods, but no repeated simulations. This fact allows for a detailed analysis with regard to data set size.

We monitor the convergence of the average weights $P(M_k \mid \mathbf{y}_{o,l})$ with increasing number of synthetic data sets $N_{d,k}$. This convergence analysis confirmed that 1000 synthetic data sets per model suffice to reach a stable result for the average weights.

4.2. Results and discussion

Fig. 5 summarizes the results of the model justifiability analysis applied to the four competing parameterizations. Fig. 5a shows the model confusion matrix as obtained when using a single pumping test in the inversion, Fig. 5b and c shows the model confusion matrices based on three and six pumping tests, respectively. The colors of the matrix columns are chosen according to the data-generating model, and the color intensity increases with increasing model weight.

The simplest model (homogeneous medium) obtains a significantly higher weight than the competing parameterizations when it generated the data (first column of the matrices), even if only using a single pumping test. A perfect justifiability with a model weight of 100% is achieved when using six pumping tests. This means that its complexity is sufficiently supported even with the smallest data set, and perfectly supported by the largest data set. Following the same argumentation, the slightly more complex zonated model is also justified in all configurations, but with less confidence than the simpler homogeneous model.

The interpolated model can hardly be self-identified when using a single pumping test, because all four models receive a posterior model weight close to their prior weight of 25%. This means that either the complexity of the interpolated model is not yet justified given 35 observations, and/or the models produce very similar predictions such that they cannot be discriminated. Since we have shown that both of the simpler models can be self-identified rather clearly based on one pumping test, the indecisive ranking between the interpolated model, the zonated model, and the homogeneous model indicates that the complexity of the interpolated model is not yet supported by the smallest data set. When including more data, the interpolated model can be self-identified and justified with increasing confidence. However, it never reaches a model weight of more than 50%, which means that there is never an "absolute majority" in favor of justifiability for this model.

The most complex model (geostatistical random fields) cannot be justified with these measurement configurations. Including more pumping test data yields a clearer decision as indicated by the increasing deviation of the posterior model weights from the prior weights. However, this decision is in favor of the less complex, but similarly structured interpolated model, and not in favor of the geostatistical model, even though it is now the one that in fact generated the data. Thus, using six pumping tests with 35 observations each does not yet suffice to justify its high level of complexity. Increasing the amount of pumping test data further is expected to lead to an even clearer decision in favor of the more parsimonious model, until a breaking point is reached at which the true underlying complexity is finally justified in light of the available data. Due to correlations in the observed data, this breaking point might require a too large amount of calibration data which cannot be accomplished in real applications.

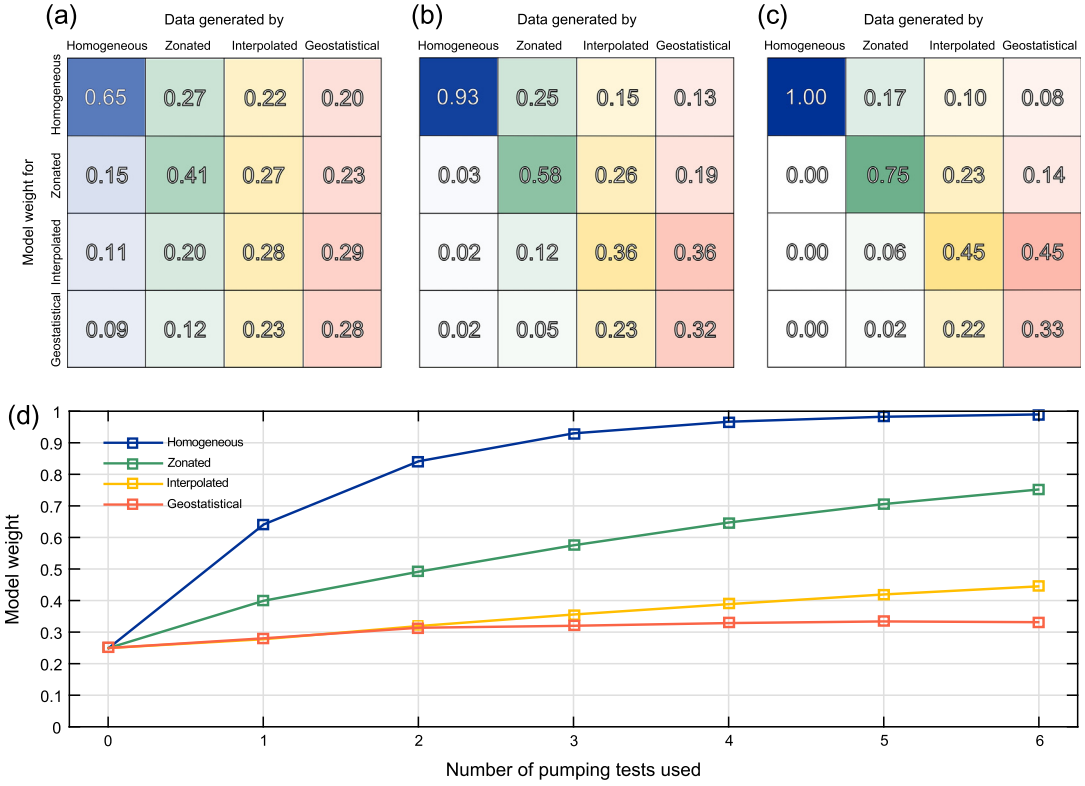


Fig. 5. Justifiability of the four models when using an increasing amount of pumping test data. (a–c) Model confusion matrices as obtained when including one, three, and six cross-hole pumping tests, respectively. (d) Average model weight for the data-generating model over increasing amount of used cross-hole pumping tests.

In general, it can be observed that with more data, any model can be self-identified with higher confidence, because the values on the main diagonal increase monotonically with increasing data set size. This is in line with BMA theory, which claims unique identification of the true model (if it is part of the considered model set) in the limit of infinite data set size. Further, within each model confusion matrix, the color intensity of the main diagonal entry decreases from the top left to the bottom right, because the higher-complexity models need more data to be justified than the lower-complexity models. The fact that the simplest model always obtains the highest weight on the main diagonal demonstrates that BMA is able to identify the true underlying model, even if it is of low complexity and a large data set is used. This confirms that BMA is not prone to over-fitting the data, as opposed to, e.g., model selection with the AIC (Burnham and Anderson, 2004).

The weights obtained by each model when it actually generated the data are plotted over increasing data set size in Fig. 5d. This graph shows that, for the simplest model, perfect justification through a model weight of 100% is approached in an asymptotic behavior. This asymptotic phase is not yet reached in the case of the zonated or the interpolated model, indicating that their level of complexity still requires a significantly larger amount of data to be perfectly justified. The model weight for the most complex model stagnates at 33% in the range of data set sizes investigated here. As explained before, we expect that a drastically larger data set would be required to justify the geostatistical model.

5. Comparison of the four parameterizations based on observed data

In this section, we present the results of inverting the actually observed hydraulic tomography data. Based on these data, we perform the standard BMA analysis to rank the four hydraulic conductivity parameterizations according to their plausibility in light of the observed data. We will interpret the obtained ranking based on our findings from the model justifiability analysis (previous section). For details on the numerical implementation, we refer to Section 3.5.

5.1. Results

5.1.1. Tomograms

By inverting the observed drawdown data from all six pumping tests, a so-called hydraulic conductivity tomogram is obtained for each parameterization. It is determined as the posterior mean of hydraulic conductivity, denoted as $E[\mathbf{u}_k \mid M_k, \mathbf{y}_o]$. Such tomograms are computed based on all four parameterizations. Fig. 6 shows these tomograms in the left column. The right column shows their respective uncertainty, determined as the posterior variance of hydraulic conductivity, $V[\mathbf{u}_k \mid M_k, \mathbf{y}_o]$.

The tomogram most rich in contrast is the one obtained from the zonated model. Because of its low complexity, its parameter values have been constrained efficiently by the inversion. This is apparent from the comparatively low posterior variance in the hydraulic conductivity estimate. The two more complex, geostatistics-based models yield a smoother hydraulic conductivity estimate in combination with a partly higher uncertainty. Among those two models, the interpolated model yields a lower uncertainty. This is due to the fact that the interpolation between the pilot points via kriging is deterministic, such that there is generally a lower uncertainty between pilot points than in the fully geostatistical field. The striped pattern in its posterior variance (Fig. 6g) is a well-known artifact of localizing the inversion to pilot points.

5.1.2. Model performance in inversion

We assess the success of the inversion for each parameterization as root mean square error (RMSE) produced by the posterior mean prediction with respect to the data used in the inversion:

$$\mathrm{RMSE}_k = \sqrt{\frac{1}{N_s}\sum \left(E[\mathbf{y} \mid \mathbf{y}_o, M_k] - \mathbf{y}_o\right)^2}. \tag{7}$$

N_s is the size of the calibration data set $\mathbf{y}_o$. $E[\mathbf{y} \mid \mathbf{y}_o, M_k]$ denotes the posterior mean prediction of model M_k. The performances of each parameterization when using data from one to six pumping tests in the inversion are listed in Table 1. The RMSE values for individual combinations of pumping tests are averaged to obtain one representative value per data set size.

The homogeneous model performs worst among the alternative parameterizations. It produces the highest error for all data set sizes. The two geostatistics-based models produce much lower errors, with the geostatistical model leading to only very minor improvements over the less complex interpolated model. A larger improvement could be expected when using an even finer resolution, i.e., a higher number of grid cells per correlation length. The zonated model performs slightly worse than the geostatistics-based models, but much better than the homogeneous model. It is worth noting that all models show a slight increase in the error when increasing the data set size from one to three pumping tests. When including even more data, however, the error reduces again, with the exception of the homogeneous model. The error reduction is due to the stronger conditioning effect in the inversion which increases the skill of the more flexible models.

5.1.3. Model ranking

We perform the BMA analysis for the data set variants described in Section 4.1, i.e., for different data set sizes (ranging from one to six pumping tests) and different combinations of pumping ports. We then average the obtained model weights over all data set variants of a specific size. The average model weights as a function of data set size (number of pumping tests used) are shown in Fig. 7a.

It can be observed that, starting from a uniform prior distribution of model weights (using zero pumping tests for calibration), model choice becomes increasingly clear with increasing number of pumping test data used in the inversion. The zonated model is clearly favored over the competing models in all data set sizes. It obtains a weight of more than 50% when using at least two pumping tests for inversion, and reaches a weight of 76% when using six pumping tests. While the weight of the zonated model increases over increasing data set size, the weights of the homogeneous and the geostatistical model decrease to less than 1% and 6%, respectively. The interpolated model shows a relatively stable weight of about 16%. When increasing the data set size from four pumping tests to six pumping tests, there is a very slight increase in model weight for the interpolated model from 15.5% to 17.2% which could be related to a slow shift in justified complexity. When using the largest data set considered here, the zonated model is clearly the most adequate parameterization. The two geostatistics-based models still obtain significant weights that justify keeping those models in the set, whereas the homogeneous model could be rejected based on a weight of 0.3%. In general, it should be checked prior to discarding a model from the set whether this model still contributes significantly to the

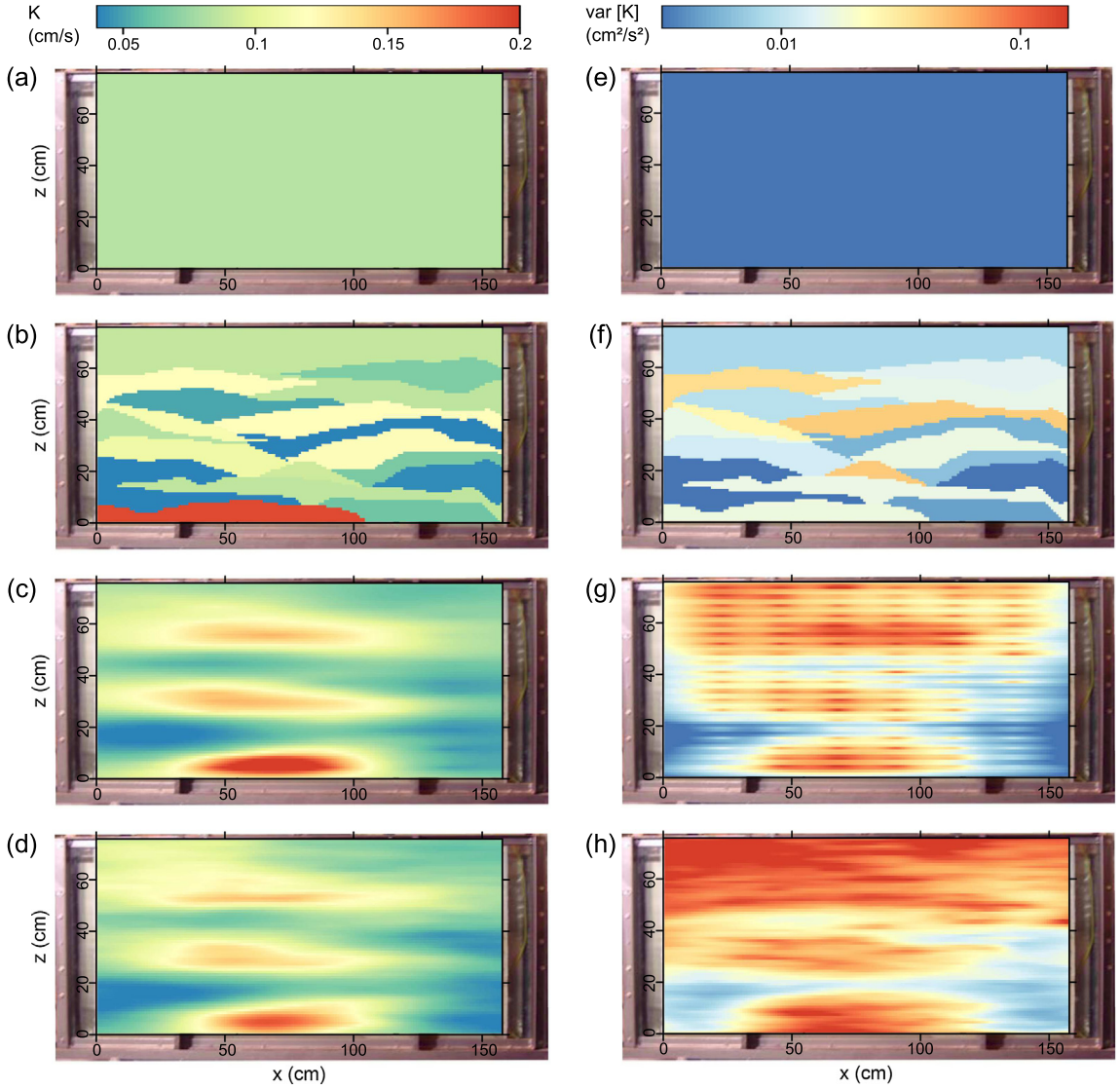


Fig. 6. Tomograms (left column) and their uncertainty (posterior parameter variance, right column) obtained from the four different parameterizations when using six pumping tests in ports 44, 47, 32, 35, 14, and 17. (a, e) Homogeneous model; (b, f) zonated model; (c, g) interpolated model; (d, h) geostatistical model.

Table 1
Average RMSE in drawdown predictions (cm) as obtained by the posterior mean of each parameterization.

Model	Number of pumping tests included in inversion					
	1	2	3	4	5	6
Homogeneous	0.35	0.40	0.41	0.42	0.42	0.42
Zonated	0.22	0.23	0.23	0.23	0.22	0.21
Interpolated	0.19	0.20	0.20	0.19	0.18	0.17
Geostatistical	0.19	0.19	0.19	0.18	0.18	0.17

between-variance part of the overall prediction variance. Previous studies have shown that this can be the case despite a very small model weight (Rojas et al., 2010; Wöhling et al., 2015).

5.1.4. Model performance in validation

We finally want to assess the performance and plausibility of the four models after conditioning on the data. Several approaches are available for validation, such as collecting additional data from core samples, single-hole pumping tests or flux measurements. Studies by Liu et al. (2007) and Illman et al. (2007) have indicated that, for large-scale groundwater models, validation with independent cross-hole pumping tests is most appropriate, because they provide more integral information about the adequacy of the models in large regions of the model domain.

We also include the BMA predictions in the validation assessment to find out whether the combination of the predictive distributions (Eq. (1)) is able to outperform the individual models in this specific case.

We compare the individual and the BMA-based model performance in predicting the drawdown distribution in the sandbox during pumping tests in ports 39, 40, 15, and 16 (see Fig. 2). These pumping test data were not used for calibration and thus allow for an independent validation of the models.

We evaluate three different performance metrics to cover aspects of accuracy, precision, and predictive coverage. To assess the accuracy of the models, we again determine the RMSE (Eq. (7)), now with respect to drawdown predictions $\boldsymbol{\varphi}$ for the validation pumping tests. For the model-weighted average, $E[\boldsymbol{\varphi} \mid \mathbf{y}_o, M_k]$ is replaced by $E[\boldsymbol{\varphi} \mid \mathbf{y}_o]$ (Eq. (2)).

As a measure of precision, we determine the average posterior variance in the predicted values of the individual models $V[\boldsymbol{\varphi} \mid \mathbf{y}_o, M_k]$ and of the model-weighted average $V[\boldsymbol{\varphi} \mid \mathbf{y}_o]$ (Eq. (3)).

We determine the predictive coverage (Hoeting et al., 1999) as the fraction of observed data points that fall within the 90%

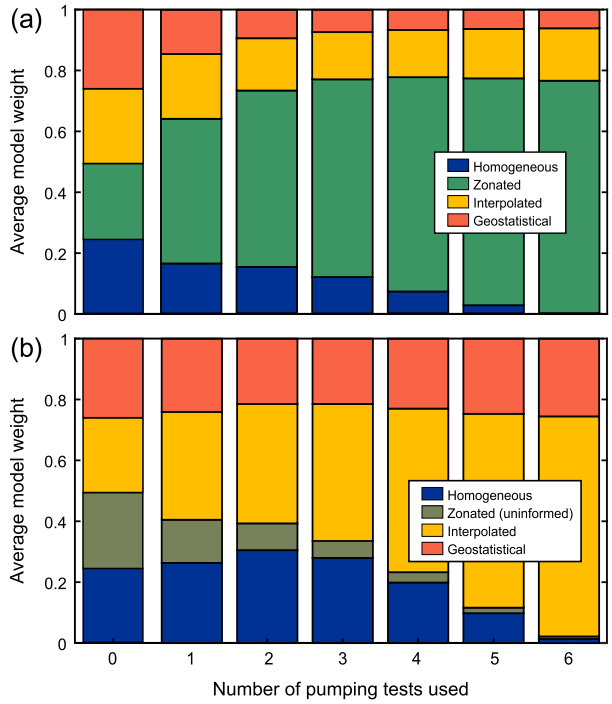


Fig. 7. Model ranking results when using an increasing amount of observed pumping test data. (a) Model weights obtained for the four parameterizations, (b) model weights obtained when replacing the geologically-informed zonation with an uninformed zonation.

Bayesian credible intervals predicted by the individual models and the model-weighted average.

The resulting statistics for the four parameterizations and their BMA-weighted average are listed in Table 2. In terms of accuracy, both geostatistics-based models yield the best results. The zonated model performs slightly worse, but still much better than the homogeneous model. The model-weighted average reflects a compromise solution between the three more complex models.

In terms of precision, the homogeneous model yields an overly confident prediction, while the geostatistics-based models show the largest uncertainty in their predictions. The model-weighted ensemble spread is larger than the individual model uncertainties.

Finally, the predictive coverage is best (but still too low) in the case of the interpolated model. The predictive coverage values achieved by the zonated model, the geostatistical model, and the BMA ensemble are slightly worse, while the homogeneous model shows a really poor coverage. Predictive coverage is especially low for all models in the case of pumping in port 15. From the layering in the sandbox (Fig. 2), it is apparent that this pumping port is located directly at a layer boundary. The zonated model is best able to reproduce the drawdown distribution there, with the highest predictive coverage and the lowest RMSE among all four models. The lower ports 39 and 40 are both located in a layer which has not been pumped in the calibration setup. These pumping tests thus provide new information on the heterogeneity in hydraulic conductivity, which is not yet fully resolved by the calibrated parameterizations. This is also apparent from the relatively high posterior variance of the tomograms (right column of Fig. 6) in the vicinity of these two ports.

Table 2
Performance of the four models and the BMA-weighted prediction in validation setup, predicting drawdown distributions induced by pumping in ports 39, 40, 15, and 16.

Model	RMSE (cm)	Standard deviation (cm)	Predictive coverage[1] (%)
Homogeneous	0.52	0.06	18
Zonated	0.37	0.20	70
Interpolated	0.33	0.22	74
Geostatistical	0.34	0.23	70
BMA	0.36	0.27	68

[1] Based on 90% prediction intervals.

5.2. Discussion

The results of the BMA analysis based on the observed data have shown that the zonated model ranks first in all measurement configurations tested here. There are two possible reasons for a model to win the contest: either it scores because of its simplicity, or because of its fit to the data (or because of both, which characterizes a most plausible model in the spirit of Occam's razor). To investigate which of these reasons are in effect here, we pursue three complimentary lines of discussion.

First, we have a more detailed look at model performance. The zonated model with its structure derived from the visible layering of the sandbox produces an acceptably small error in all data set sizes. There is no sign of over-fitting the data, since the RMSE stays almost constant over increasing data set size. The much more parsimonious homogeneous model performs worst among the alternative parameterizations. Thus, its poor BMA model ranking result clearly stems from a significant mismatch between predicted and observed drawdown, which cannot be compensated by the minimum complexity of the model. The smallest errors are produced by the much more flexible geostatistics-based models. The most complex model, however, cannot improve significantly on its less complex interpolated variant. This indicates that the additional complexity cannot be exploited in the inversion to achieve a better fit. Thus, the zonated model seems to provide sufficient flexibility (but no more than that) at the right places to score in both disciplines, goodness-of-fit and parsimony.

As second line of discussion, we compare the model weight distribution based on the observed data with the model weight distribution obtained in the justifiability analysis (Section 4.2). From the justifiability analysis, we know that the zonated model can be well identified with a weight of 75% if it truly is the data-generating model. When using the actually observed data, it obtains practically the same weight (76%) in the same experimental setup. Thus, the zonated model lives up to its full potential, which is a strong indication that it is in fact close to the observed system response.

To cover this interpretation, we hypothesize for a moment that this conclusion is not true, and that the underlying system is actually closer to the more complex interpolated model. For that case, the justifiability analysis predicts that the weight for the interpolated model should be roughly twice the weight of the zonated model. This expectation, however, is not reflected in the ranking based on the observed data. Apparently, the interpolated model does not live up to its potential determined in the justifiability analysis. We attribute this to a significant mismatch with regard to the observed data, and disregard our alternative hypothesis. Again, we find that the zonated model is not only favored due to its simplicity.

We verify this with a third test. We introduce another competitor into the model set which is similarly parsimonious as the zonated model, but uses an apparently wrong (i.e., geologically uninformed) structure. We define this additional model as zonated parameterization with 24 rectangular zones, distributed regularly and equispaced over the model domain (4 columns $\times$ 6 rows). While this zonation is at least oriented horizontally along the main axis of the visible layering in the sandbox, it does not benefit from our knowledge about the actual geometry of the layers. The model weights obtained when replacing the geologically informed

zonation with the geologically uninformed variant are shown in Fig. 7b. The uninformed variant is outperformed by all three alternative models for all data set sizes. When using more than three pumping tests in the inversion, the interpolated model achieves an absolute majority of the weights, and the weight of the uninformed zonation model is reduced to less than 5%. Here, even the homogeneous model receives a weight higher than its prior model weight when using one or two pumping tests in the inversion. This reflects that the interpolated model still lacks justifiability under these experimental setups as demonstrated in the justifiability analysis. The uninformed variant of the zonated model provides a degree of flexibility similar to the informed variant, but this flexibility is allowed at the wrong places. We again conclude that the zonated model with its structure derived from the visible packing pattern scores because it combines low complexity with good performance.

To sum up, we have thoroughly investigated the factors contributing to the success of the geologically-informed zonated model in the ranking and conclude that it meets both criteria that make up the Bayesian tradeoff, goodness-of-fit *and* parsimony (i.e., relatively small variance in its predictions). This model therefore wins the competition through the eyes of Occam's razor. On the other hand, we have illustrated that a zonation-based model which is not informed by the apparent layering will not be able to compete with either the simpler homogeneous medium approach or the more flexible geostatistics-based approaches.

Finally, we have compared the skills of the individual models and the model-weighted average in a validation setup. Results have shown that the zonated model represents a compromise solution between accuracy, prediction uncertainty and predictive coverage. In this specific experimental setup, the winner of the model competition even outperforms the BMA-weighted average.

6. Summary and conclusions

In this study, we address the question which complexity of a model is justified given a specific amount of calibration data. Model complexity as such can be defined in various ways, ranging from simple parameter counting over factor analysis to concepts that also account for data-parameter sensitivity. We propose to use Bayesian model averaging (BMA) to measure complexity and to identify the justified degree of complexity, because BMA is a multi-model approach that implicitly follows the principle of parsimony and avoids over-fitting of the data. BMA determines the optimal tradeoff between goodness-of-fit and complexity based on Bayes' theorem. We lay a special focus on the characteristics of this tradeoff by investigating the shift in justified complexity with an increasing amount of calibration data.

We suggest a two-step BMA procedure to identify the optimal balance between data requirements and model complexity: First, a *model justifiability analysis* is performed in a synthetic setup to determine the maximum level of complexity that could possibly be justified with the given type and amount of experimental data. This analysis allows us to assess the complexity aspect of the BMA tradeoff isolated from the performance aspect. Second, the standard BMA analysis is performed to assess the *adequacy* of each model based on the observed data. Results are then interpreted in light of the findings from the model justifiability analysis. This two-step procedure answers the question, whether the model ranking first in the BMA analysis is really the best choice given the current set of models, or if it is only optimal given the currently too limited amount of data which does not justify a more complex model among the set of considered models, although the more complex one would actually be closer to the observed system response. Further, the justifiability analysis can reveal whether two models that receive the same posterior weight in the standard BMA procedure are actually very similar in their predictions or whether the same weight just indicates a similar overall goodness-of-fit.

A main contribution of our study is the proposed procedure for the first step, the model justifiability analysis. We want to systematically test whether, given a specific experimental setup, BMA is able to identify the respective true underlying complexity. To achieve this, we let each of the alternative models generate many random synthetic data sets, mimicking the measurement configuration of the experimental setup that produces the real data. The standard BMA analysis is then performed based on these synthetic data sets and the resulting BMA weights are averaged per data-generating model. By looping over the model set, we populate what we call a *model confusion matrix*. This matrix expresses how likely it is to identify any specific model from the set of considered models if it was actually the true one, given the current experimental setup. If a specific degree of model complexity can be self-identified with significance, we call it *justifiable* through the eyes of BMA.

We have illustrated our suggested approach with an application to groundwater model calibration via hydraulic tomography. The drawdown in a synthetic sandbox aquifer induced by several pumping tests is simulated with a groundwater model. Four alternative parameterizations of hydraulic conductivity are considered, which differ in complexity by orders of magnitude: a homogeneous model with one effective parameter, a zonated model inspired by the visible layering in the sandbox with a few zones (here: 19), a model based on interpolation between pilot points (here: 120), and a geostatistical model that allows all cells (here: 12,480) to vary randomly according to the prescribed geostatistical model.

Results of our two-step procedure have shown that, given the synthetic data from up to six pumping tests, the geostatistical model is not yet justified through the eyes of BMA. Only the lower-complexity models can be self-identified and therefore justified sufficiently well. When using the actually observed data, the zonated model with its structure derived from the visible layering of the sandbox ranks first in all investigated experimental setups. The zonated model wins the model competition because it shows a sufficiently small degree of complexity (which results in a sufficiently small predictive variance), and it shows a good quality of fit. Thus, it scores with respect to both counterparts of the Bayesian tradeoff. The zonated model is therefore the most adequate representation of the true system (at least on the basis of drawdown data) out of the considered models. This finding is in line with the expectation that sharp layer boundaries can be well approximated by the zonated model with its structure derived from the visible layering. The geostatistics-based models also have their merits according to non-negligible model weights, probably because they are flexible enough to image mixing effects at those boundaries. It is important to note, however, that a zonation which is not inspired by the true layering will not be able to compete with geostatistics-based approaches. We have finally assessed the performance of the different parameterizations in a validation setup. The geologically-informed zonation represents a compromise solution between accuracy, prediction uncertainty and predictive coverage.

In summary, our results suggest that aquifer characterization via hydraulic tomography does not necessarily justify a geostatistical description of aquifer heterogeneity. Instead, geology-based zonation might be a more robust choice, but only if reliable information about the layering is available. In practice, such a zonated model could even be further equipped with prior knowledge by prescribing correlations between layers/zones of similar material. Pursuing the idea of combining the "best of two worlds" beyond a linear combination of predictive distributions as done by BMA,

a combination of a zonated model with a geostatistical description of small-scale heterogeneities within the zones (see, e.g., Fienen et al., 2008) could prove useful and most plausible in practical applications. Further, multi-point geostatistics or multi-modal marginal distributions might have larger geological reality, but they are also conceptually more complex. Future research should target the question whether such models would objectively be favored in a BMA analysis given a realistically available amount of data, and under which data types they would be favored. Since the main difference in model behavior would be connectivity, we expect that choosing such a type of model is more beneficial when predicting solute transport than when predicting flow conditions.

The proposed model justifiability analysis is a very general upgrade of the BMA procedure, as it is applicable to any type of models and data. Given a sufficient budget of computation time, the two-step BMA procedure is expected to facilitate the interpretation of the resulting model ranking tremendously, especially in field-scale applications. This is open to be tested in future studies.

Acknowledgments

The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the International Research Training Group "Integrated Hydrosystem Modelling" (IRTG 1829) at the University of Tübingen and within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart. Additional support for this study was provided to WAI by the Natural Resources and Engineering Council of Canada (NSERC). The authors also thank Ming Ye and the anonymous reviewer for their constructive comments on the manuscript. The data used in this study can be requested from Walter A. Illman (willman@uwaterloo.ca).

References

Ajami, N.K., Gu, C., 2010. Complexity in microbial metabolic processes in soil nitrogen modeling: a case for model averaging. Stoch. Environ. Res. Risk Assess. 24 (6), 831–844. http://dx.doi.org/10.1007/s00477-010-0381-4.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. pp. 367–281.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference – understanding AIC and BIC in model selection. Sociol. Meth. Res. 33 (2), 261–304. http://dx.doi.org/10.1177/0049124104268644.

Butler, J.J., McElwee, C.D., Bohling, G.C., 1999. Pumping tests in networks of multilevel sampling wells: motivation and methodology. Water Resour. Res. 35 (11), 3553–3560. http://dx.doi.org/10.1029/1999wr900231.

Cardiff, M., Barrash, W., 2011. 3-D transient hydraulic tomography in unconfined aquifers with fast drainage response. Water Resour. Res. 47 (12). http://dx.doi.org/10.1029/2010wr010367.

Cardiff, M., Barrash, W., Kitanidis, P.K., Malama, B., Revil, A., Straface, S., Rizzo, E., 2009. A potential-based inversion of unconfined steady-state hydraulic tomography. Ground Water 47 (2), 259–270. http://dx.doi.org/10.1111/j.1745-6584.2008.00541.x.

Castagna, M., Bellin, A., 2009. A Bayesian approach for inversion of hydraulic tomographic data. Water Resour. Res. 45 (4). http://dx.doi.org/10.1029/2008wr007078. Artn W04410.

Draper, D., 1995. Assessment and propagation of model uncertainty. J. Roy. Stat. Soc. Ser. B-Methodol. 57 (1), 45–97.

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. J. Hydrol. 517, 105–119. http://dx.doi.org/10.1016/j.jhydrol.2014.05.027.

Elsheikh, A.H., Wheeler, M.F., Hoteit, I., 2013. Nested sampling algorithm for subsurface flow model selection, uncertainty quantification, and nonlinear calibration. Water Resour. Res. 49 (12), 8383–8399. http://dx.doi.org/10.1002/2012wr013406.

Fienen, M.N., Clemo, T., Kitanidis, P.K., 2008. An interactive Bayesian geostatistical inverse protocol for hydraulic tomography. Water Resour. Res. 44 (12). http://dx.doi.org/10.1029/2007wr006730.

Fienen, M., Hunt, R., Krabbenhoft, D., Clemo, T., 2009. Obtaining parsimonious hydraulic conductivity fields using head and transport observations: a Bayesian geostatistical parameter estimation approach. Water Resour. Res. 45 (8).

Foglia, L., Mehl, S.W., Hill, M.C., Burlando, P., 2013. Evaluating model structure adequacy: the case of the Maggia Valley groundwater system, southern Switzerland. Water Resour. Res. 49 (1), 260–282. http://dx.doi.org/10.1029/2011wr011779.

George, E.I., 2010. Dilution priors: compensating for model space redundancy. Inst. Math. Stat., 158–165

Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. Novel approach to nonlinear non-Gaussian Bayesian state estimation. IEE Proc.-F Radar Signal Process. 140 (2), 107–113.

Gottlieb, J., Dietrich, P., 1995. Identification of the permeability distribution in soil by hydraulic tomography. Inverse Prob. 11 (2), 353–360. http://dx.doi.org/10.1088/0266-5611/11/2/005.

Gull, S.F., 1988. Bayesian inductive inference and maximum entropy. In: Maximum Entropy and Bayesian Methods in Science and Engineering, vols. 31–32, pp. 53–74.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14 (4), 382–401.

Hubbard, S.S., Rubin, Y., 2000. Hydrogeological parameter estimation using geophysical data: a review of selected techniques. J. Contam. Hydrol. 45 (1–2), 3–34. http://dx.doi.org/10.1016/S0169-7722(00)00117-0.

Huelsenbeck, J.P., Larget, B., Alfaro, M.E., 2004. Bayesian phylogenetic model selection using reversible jump Markov Chain Monte Carlo. Mol. Biol. Evol. 21 (6), 1123–1133. http://dx.doi.org/10.1093/molbev/msh123.

Illman, W.A., Liu, X.Y., Craig, A., 2007. Steady-state hydraulic tomography in a laboratory aquifer with deterministic heterogeneity: multi-method and multiscale validation of hydraulic conductivity tomograms. J. Hydrol. 341 (3–4), 222–234. http://dx.doi.org/10.1016/j.jhydrol.2007.05.011.

Illman, W.A., Zhu, J., Craig, A.J., Yin, D., 2010. Comparison of aquifer characterization approaches through steady state groundwater model validation: a controlled laboratory sandbox study. Water Resour. Res. 46 (4). http://dx.doi.org/10.1029/2009wr007745.

Illman, W.A., Berg, S.J., Zhao, Z., 2015. Should hydraulic tomography data be interpreted using geostatistical inverse modeling? A laboratory sandbox investigation. Water Resour. Res. 51 (5), 3219–3237. http://dx.doi.org/10.1002/2014WR016552.

Jeffreys, H., 1939. Theory of Probability. Oxford University Press, chapter 5.

Kaiser, H.F., 1960. The application of electronic computers to factor analysis. Educ. Psychol. Meas. 20, 141–151.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90 (430), 773–795. http://dx.doi.org/10.2307/2291091.

Kitanidis, P.K., 1995. Quasi-linear geostatistical theory for inversing. Water Resour. Res. 31 (10), 2411–2419. http://dx.doi.org/10.1029/95WR01945.

Kitanidis, P.K., 1997. Introduction to Geostatistics: Applications in Hydrogeology. Cambridge University Press.

Li, W., Nowak, W., Cirpka, O.A., 2005. Geostatistical inverse modeling of transient pumping tests using temporal moments of drawdown. Water Resour. Res. 41 (8). http://dx.doi.org/10.1029/2004wr003874. Artn W08403.

Li, W., Englert, A., Cirpka, O.A., Vanderborght, J., Vereecken, H., 2007. Two-dimensional characterization of hydraulic heterogeneity by multiple pumping tests. Water Resour. Res. 43 (4). http://dx.doi.org/10.1029/2006wr005333.

Liu, J.S., 2008. Monte Carlo Strategies in Scientific Computing. Springer, New York.

Liu, X., Illman, W.A., Craig, A.J., Zhu, J., Yeh, T.C.J., 2007. Laboratory sandbox validation of transient hydraulic tomography. Water Resour. Res. 43 (5). http://dx.doi.org/10.1029/2006wr005144.

Morales-Casique, E., Neuman, S.P., Vesselinov, V.V., 2010. Maximum likelihood Bayesian averaging of airflow models in unsaturated fractured tuff using Occam and variance windows. Stoch. Environ. Res. Risk Assess. 24 (6), 863–880. http://dx.doi.org/10.1007/s00477-010-0383-2.

Najafi, M.R., Moradkhani, H., Jung, I.W., 2011. Assessing the uncertainties of hydrologic model selection in climate change impact studies. Hydrol. Process. 25 (18), 2814–2826. http://dx.doi.org/10.1002/Hyp.8043.

Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stoch. Environ. Res. Risk Assess. 17 (5), 291–305. http://dx.doi.org/10.1007/s00477-003-0151-7.

Neuman, S.P., Xue, L., Ye, M., Lu, D., 2012. Bayesian analysis of data-worth considering model and parameter uncertainties. Adv. Water Resour. 36, 75–85. http://dx.doi.org/10.1016/j.advwatres.2011.02.007.

Nowak, W., Schwede, R.L., Cirpka, O.A., Neuweiler, I., 2008. Probability density functions of hydraulic head and velocity in three-dimensional heterogeneous porous media. Water Resour. Res. 44 (8). http://dx.doi.org/10.1029/2007wr006383. Artn W08452.

Nowak, W., de Barros, F.P.J., Rubin, Y., 2010. Bayesian geostatistical design: task-driven optimal site investigation when the geostatistical model is uncertain. Water Resour. Res. 46 (3). http://dx.doi.org/10.1029/2009wr008312. Artn W03535.

Nowak, W., Rubin, Y., de Barros, F.P.J., 2012. A hypothesis-driven approach to optimize field campaigns. Water Resour. Res. 48 (6). http://dx.doi.org/10.1029/2011wr011016.

Raftery, A.E., 1995. Bayesian model selection in social research. Sociol. Methodol. 1995 (25), 111–163. http://dx.doi.org/10.2307/271063.

RamaRao, B.S., LaVenue, A.M., de Marsily, G., Marietta, M.G., 1995. Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields. 1. Theory and computational experiments. Water Resour. Res. 31 (3), 475–493. http://dx.doi.org/10.1029/94wr02258.

Refsgaard, J.C., Christensen, S., Sonnenborg, T.O., Seifert, D., Hojberg, A.L., Troldborg, L., 2012. Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. Adv. Water Resour. 36, 36–50. http://dx.doi.org/10.1016/j.advwatres.2011.04.006.

Rojas, R., Feyen, L., Dassargues, A., 2008. Conceptual model uncertainty in groundwater modeling: combining generalized likelihood uncertainty estimation and Bayesian model averaging. Water Resour. Res. 44 (12). http://dx.doi.org/10.1029/2008wr006908.

Rojas, R., Batelaan, O., Feyen, L., Dassargues, A., 2010. Assessment of conceptual model uncertainty for the regional aquifer Pampa del Tamarugal – North Chile. Hydrol. Earth Syst. Sci. 14 (2), 171–192.

Schöniger, A., Nowak, W., Franssen, H.J.H., 2012. Parameter estimation by ensemble Kalman filters with transformed data: approach and application to hydraulic tomography. Water Resour. Res. 48 (4). http://dx.doi.org/10.1029/2011wr010462. Artn W04502.

Schöniger, A., Wöhling, T., Samaniego, L., Nowak, W., 2014. Model selection on solid ground: rigorous comparison of nine ways to evaluate Bayesian model evidence. Water Resour. Res. 50 (12), 9484–9513. http://dx.doi.org/10.1002/2014WR016062.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464. http://dx.doi.org/10.1214/aos/1176344136.

Seifert, D., Sonnenborg, T.O., Refsgaard, J.C., Hojberg, A.L., Troldborg, L., 2012. Assessment of hydrological model predictive ability given multiple conceptual geological models. Water Resour. Res. 48 (6). http://dx.doi.org/10.1029/2011wr011149, doi: Artn W06503.

Singh, A., Mishra, S., Ruskauff, G., 2010. Model averaging techniques for quantifying conceptual model uncertainty. Ground Water 48 (5), 701–715. http://dx.doi.org/10.1111/j.1745-6584.2009.00642.x.

Straface, S., Yeh, T.C.J., Zhu, J., Troisi, S., Lee, C.H., 2007. Sequential aquifer tests at a well field, Montalto Uffugo Scalo, Italy. Water Resour. Res. 43 (7). http://dx.doi.org/10.1029/2006wr005287.

Sudicky, E.A., 1986. A natural gradient experiment on solute transport in a sand aquifer: spatial variability of hydraulic conductivity and its role in the dispersion process. Water Resour. Res. 22 (13), 2069–2082. http://dx.doi.org/10.1029/WR022i013p02069.

Sudicky, E.A., Illman, W.A., Goltz, I.K., Adams, J.J., McLaren, R.G., 2010. Heterogeneity in hydraulic conductivity and its role on the macroscale transport of a solute plume: from measurements to a practical application of stochastic flow and transport theory. Water Resour. Res. 46. http://dx.doi.org/10.1029/2008wr007558.

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 Resour. Res. 46 (12). http://dx.doi.org/10.1029/2010wr009227.

Tsai, F.T.C., Elshall, A.S., 2013. Hierarchical Bayesian model averaging for hydrostratigraphic modeling: uncertainty segregation and comparative evaluation. Water Resour. Res. 49 (9), 5520–5536. http://dx.doi.org/10.1002/Wrcr.20428.

Tsai, F.T.-C., Li, X.B., 2008. Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window. Water Resour. Res. 44 (9). http://dx.doi.org/10.1029/2007wr006576.

Vesselinov, V.V., Neuman, S.P., Illman, W.A., 2001. Three-dimensional numerical inversion of pneumatic cross-hole tests in unsaturated fractured tuff 2. Equivalent parameters, high-resolution stochastic imaging and scale effects. Water Resour. Res. 37 (12), 3019–3041. http://dx.doi.org/10.1029/2000wr000135.

Wöhling, T., Schöniger, A., Gayler, S., Nowak, W., 2015. Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction. Water Resour. Res. 51 (4), 2825–2846. http://dx.doi.org/10.1002/2014wr016292.

Xue, L., Zhang, D., Guadagnini, A., Neuman, S.P., 2014. Multimodel Bayesian analysis of groundwater data worth. Water Resour. Res. 50 (11), 8481–8496. http://dx.doi.org/10.1002/2014wr015503.

Ye, M., Neuman, S.P., Meyer, P.D., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour. Res. 40 (5). http://dx.doi.org/10.1029/2003wr002557.

Ye, M., Meyer, P.D., Neuman, S.P., 2008. On model selection criteria in multimodel analysis. Water Resour. Res. 44 (3). http://dx.doi.org/10.1029/2008wr006803.

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. Ground Water 48 (5), 716–728. http://dx.doi.org/10.1111/j.1745-6584.2009.00633.x.

Yeh, T.C.J., Liu, S.Y., 2000. Hydraulic tomography: development of a new aquifer test method. Water Resour. Res. 36 (8), 2095–2105. http://dx.doi.org/10.1029/2000wr900114.

Yeh, W.W.G., Yoon, Y.S., 1981. Aquifer parameter-identification with optimum dimension in parameterization. Water Resour. Res. 17 (3), 664–672. http://dx.doi.org/10.1029/Wr017i003p00664.

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2015WR016918

A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking

Anneli Schöniger[1], Thomas Wöhling[2,3], and Wolfgang Nowak[4]

[1]Center for Applied Geoscience, University of Tübingen, Tübingen, Germany, [2]Water and Earth System Science (WESS) Competence Cluster, University of Tübingen, Tübingen, Germany, [3]Lincoln Environmental Research, Lincoln Agritech Limited, Hamilton, New Zealand, [4]Institute for Modelling Hydraulic and Environmental Systems (LS[3])/SimTech, University of Stuttgart, Stuttgart, Germany

Key Points:
- Uncertainty in model input or output data induces uncertainty in model weights
- Weighting uncertainty can compromise the confidence in model ranking
- We propose a statistical concept to account for weighting uncertainty in BMA

Correspondence to:
A. Schöniger,
anneli.schoeniger@uni-tuebingen.de

Citation:
Schöniger, A., T. Wöhling, and W. Nowak (2015), A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking, *Water Resour. Res.*, *51*, 7524–7546, doi:10.1002/2015WR016918.

Received 14 JAN 2015
Accepted 21 AUG 2015
Accepted article online 28 AUG 2015
Published online 13 SEP 2015

Abstract Bayesian model averaging (BMA) ranks the plausibility of alternative conceptual models according to Bayes' theorem. A prior belief about each model's adequacy is updated to a posterior model probability based on the skill to reproduce observed data and on the principle of parsimony. The posterior model probabilities are then used as model weights for model ranking, selection, or averaging. Despite the statistically rigorous BMA procedure, model weights can become uncertain quantities due to measurement noise in the calibration data set or due to uncertainty in model input. Uncertain weights may in turn compromise the reliability of BMA results. We present a new statistical concept to investigate this weighting uncertainty, and thus, to assess the significance of model weights and the confidence in model ranking. Our concept is to resample the uncertain input or output data and then to analyze the induced variability in model weights. In the special case of weighting uncertainty due to measurement noise in the calibration data set, we interpret statistics of Bayesian model evidence to assess the distance of a model's performance from the theoretical upper limit. To illustrate our suggested approach, we investigate the reliability of soil-plant model selection following up on a study by Wöhling et al. (2015). Results show that the BMA routine should be equipped with our suggested upgrade to (1) reveal the significant but otherwise undetected impact of measurement noise on model ranking results and (2) to decide whether the considered set of models should be extended with better performing alternatives.

1. Introduction

Conceptual uncertainty has been widely recognized as a main source of uncertainty in environmental model predictions [e.g., *Burnham and Anderson*, 2003; *Murphy et al.*, 2004; *Refsgaard et al.*, 2006; *Rojas et al.*, 2008; *Renard et al.*, 2010]. Especially for complex, coupled systems, reliable prognoses are hardly possible with a single-model approach. When working with a single model, only the uncertainty due to forcings and parameters of that single model can be assessed (within-model variance). The uncertainty due to the very choice of the most adequate representation of the system is neglected, leading to a severe underestimation of the overall predictive uncertainty. In an attempt to account for the uncertainty in model building, a set of plausible competing models should be considered instead. In such multimodel approaches, weights are assigned to each conceptual model based on goodness of fit and, often, on some penalty for complexity. With these weights, an averaged estimate can be obtained (model averaging) or a specific model may be selected on an objective basis, if there is clear evidence for this model to be the most adequate one in the set (model selection). In addition to within-model variance, conceptual uncertainty in the current set of models can be quantified as between-model variance. By reporting alternative system representations, the modeler's and decision maker's confidence in model predictions is significantly increased and the potential bias in modeling is reduced.

Bayesian model averaging (BMA) [*Draper*, 1995; *Hoeting et al.*, 1999] is a formal statistical approach that handles multiple models. It ranks alternative conceptual models according to their plausibility. The approach is based on Bayes' theorem, which updates a prior belief about the adequacy of each model in the set to a posterior probability by judging each model's performance in reproducing a calibration data set. The posterior probabilities are used as model weights. The individual predictive probability distributions of each model in the set are combined in a weighted average. This weighted average yields a robust estimate that typically outperforms the individual models in predictive coverage. BMA implicitly follows the principle of

parsimony or Occam's razor [*Gull*, 1988], so that the posterior model weights will reflect an optimal compromise between model complexity and goodness of fit. BMA is a very general framework in that it can be applied to any type of model or application in the same systematic procedure, yielding reproducible and objectively comparable results. This is a major step toward a more rigorous quantification of model prediction uncertainty, which strengthens the basis for management tasks and for risk assessment.

Technically, BMA involves the evaluation of Bayesian model evidence (BME), which is the likelihood of the observed data integrated over each model's parameter space. This integral typically cannot be solved analytically, and numerical solutions come at the price of high computational cost. Efficient alternatives in the form of mathematical approximations to the analytical BMA equations have therefore become popular, such as the Kashyap information criterion (KIC) [*Neuman*, 2003], the Bayesian information criterion (BIC) [*Schwarz*, 1978; *Raftery*, 1995], or the Akaike information criterion (AIC) [*Akaike*, 1973], to name the most frequently applied ones. Various studies have, however, revealed that these approximations yield contradicting results for model ranking [see, e.g., *Ye et al*., 2008; *Tsai and Li*, 2008; *Ye et al*., 2010; *Rojas et al*., 2010a; *Singh et al*., 2010; *Morales-Casique et al*., 2010; *Foglia et al*., 2013]. In a rigorous intercomparison study [*Schöniger et al*., 2014], the approximations have been benchmarked against reference solutions in both a synthetic and a real-world test case. Results have shown that these criteria differ in how strongly they penalize model complexity, and as a consequence, differ in their values for posterior model weights or even in the ranking of the models. For nonlinear models, the true weighting according to Bayes' theorem can only be reliably obtained from brute-force Monte Carlo integration.

BMA has been applied by researchers for various modeling tasks in the field of water resources. Disciplines include, but are not limited to, climate change modeling [e.g., *Najafi et al*., 2011], hydrological modeling [e.g., *Ajami et al*., 2007; *Vrugt and Robinson*, 2007; *Wöhling and Vrugt*, 2008], hydrogeological modeling [e.g., *Rojas et al*., 2008; *Poeter and Anderson*, 2005; *Ye et al*., 2010; *Neuman*, 2003; *Ye et al*., 2004], and soil-plant modeling [*Wöhling et al*., 2015].

Some recent studies have investigated and commented on the impact of different data set sizes and data types on the outcome of BMA weights [*Rojas et al*., 2010b; *Lu et al*., 2012; *Refsgaard et al*., 2012; *Xue et al*., 2014; *Wöhling et al*., 2015]. It has been found that model ranking can vary significantly with the size and composition of the data set and that the performance of the model-averaged predictions in validation clearly depends on the data chosen for calibration. These findings emphasize the importance of choosing an appropriate calibration data set. BMA evaluates to which degree the competing models agree with the observed data (to which degree the models are fit for purpose), and not to which degree they actually represent the underlying system. Acknowledging that the purpose changes with varied calibration data sets (at least in the eyes of BMA) calls for a further analysis whether the obtained weights are representative and robust for the application at hand.

Since BME and hence BMA weights are a function of the observed data, they not only depend on the data set chosen for calibration, but also on the very outcome of random measurement error for all individual data values. It clearly follows that there is an inherent uncertainty attached to the model weights, since they inevitably vary under different outcomes of measurement errors. *Carrera and Neuman* [1986] found that model ranking changed with an increase in the assumed level of measurement error, but did not investigate this fact any further. To our knowledge, there have been no studies yet that investigate the uncertainty in BMA weights due to measurement noise. We refer to this uncertainty as *weighting uncertainty*. Weighting uncertainty, once recognized and acknowledged, triggers the need to extend the BMA concept.

In the current study, we introduce such an extension. With our new statistical concept, we will assess the robustness of model weights against measurement noise and the related confidence in model ranking. We propose to investigate the robustness against measurement errors by perturbing the observed data with a data type-specific random measurement error and by analyzing the resulting variability in the obtained weights. If this variability is small, it can be concluded that the obtained weights are representative and not artifacts of the specific outcome of measurement errors. Model ranking, model selection, or model averaging based on this calibration data set can then be regarded as robust. In contrast, a large variability might indicate a strong sensitivity to the exact observed measurement values including their errors. This means that model selection should not be based on the current measurement design, since small changes in measurement values within the plausible range as defined by the assumed measurement error standard deviation could lead to a contradicting proposition for model choice. Also, model averaging of predictive distributions would be questionable on the basis of unreliable model weights. In this study, we focus on the robustness of model ranking as a direct result of uncertain model weights.

Generally speaking, we take a frequentist perspective upon the variability of BMA results by randomizing measurement error in the observed data. This enables us to quantify the inherent uncertainty caused by measurement noise, which would otherwise not become obvious from the existing BMA analysis. Frequentist properties of the posterior have been assessed in many different contexts [see, e.g., *Carlin and Louis*, 2000], but to our knowledge not yet within the BMA framework.

From such a resampling analysis, we determine distributions of BME for each model. We can then compare these distributions to a theoretically optimal distribution. The theoretically optimal distribution is defined by a perfect model that has no other sources of misfit to the data than measurement noise. With this comparison, we can assess how far off the model set is from the theoretically optimal model, similar to a performance metric proposed by *Abramowitz and Gupta* [2008]. Knowing this distance is of great value to modelers to decide whether the modeling task can be carried out satisfactorily with the models at hand, or whether more effort should be invested in improving the model set and/or extending it with better performing models.

While the impact of measurement noise in the calibration data set on the outcome of model ranking is our primary motivation to extend the BMA routine, our concept is generally able to also handle any other source of uncertainty for BMA weights that can be addressed by a resampling analysis. Such other sources of uncertainty could be, e.g., noisy measurements of model input or forcings or conceptual uncertainty in boundary conditions. Input uncertainty is especially relevant in the field of hydrology due to the large variability of precipitation in time and space. Different approaches have been proposed to account for input uncertainty, parameter uncertainty, and model structural uncertainty in an integrated manner, such as the Bayesian total error analysis (BATEA) [*Kavetski et al*., 2006; *Kuczera et al*., 2006] and the integrated Bayesian uncertainty estimator (IBUNE) [*Ajami et al*., 2007]. BATEA quantifies a full posterior predictive distribution considering the mentioned sources of uncertainty. It differs from our extended BMA approach in that it does not pursue the idea of model ranking or multimodel combination. Hence, it does not provide any insights on the impact of the different sources of uncertainty on model weights. The IBUNE approach combines multiple models into a weighted prediction, but obtains the model weights from a non-Bayesian scheme. Further, IBUNE handles input uncertainty by marginalization, as opposed to our extended BMA routine which explicitly aims at evaluating the variability in weights due to a specific source of uncertainty (be it in calibration data, forcings, or boundary conditions). We therefore see our proposed extended BMA routine as a valuable addition to the existing pool of methods for an integrated assessment of modeling uncertainty, because it elegantly combines the advantages of a fully Bayesian predictive uncertainty quantification with those of a Bayesian multimodel framework.

As main contributions of this study, we (1) propose a numerical framework to determine weighting uncertainty due to uncertain model input or output data, (2) demonstrate different options to visually assess the confidence in model ranking, and (3) provide a measure for the distance in performance of a model from the theoretical upper limit imposed by noise in the calibration data set. These contributions strengthen the basis of scientific model hypothesis testing.

The paper is organized as follows: after introducing our notation of model predictions and probability distributions in section 2.1, we present the statistical framework of BMA in section 2.2 and discuss the interpretation of BMA weights and BME values in section 2.3. We introduce our proposed extension to consider weighting uncertainty in section 2.4. Section 2.5 explains how to determine the upper limit in model performance under measurement noise in output data. In sections 2.6 and 2.7, we compare several alternatives to interpret the results of the resampling analysis. Section 3 features an application of our analysis tools to soil-plant model selection, following up on the BMA analysis performed by *Wöhling et al*. [2015]. We discuss the observed impact of weighting uncertainty on model ranking in section 4. Finally, we summarize our findings and implications in section 5.

2. Methodology

2.1. Notation of Models and Predictions

Let us consider N_m competing conceptual models M_k which produce model predictions $\boldsymbol{\varphi}_k$ as a function of $\boldsymbol{\Theta}_k$ and $\mathbf{c}$:

$$\boldsymbol{\varphi}_k = M_k^{(\varphi)}(\boldsymbol{\Theta}_k) = f^{(\varphi)}(\boldsymbol{\Theta}_k, \mathbf{c}), \tag{1}$$

with $\boldsymbol{\Theta}_k$ including nondeterministic variables such as uncertain model parameters $\mathbf{u}_k$, uncertain model input $\mathbf{v}$, natural (stochastic) noise $\mathbf{w}$ (aleatory uncertainty), and adjustable model structural errors $\boldsymbol{e}_k$ (epistemic uncertainty), depending on the modeling task and the choices of the modeler. Prior knowledge about these variables can be specified in the form of probability distributions: $p(\mathbf{u}|M_k), p(\mathbf{v}), p(\mathbf{w})$, and $p(\boldsymbol{e}|M_k)$, with $p(\cdot|\cdot)$ representing a conditional probability distribution. Note that we refer to model input in a broader sense, including time-variant or constant forcings and boundary conditions. Model predictions $\boldsymbol{\varphi}_k$ are further a function of deterministic variables $\mathbf{c}$, such as fixed input values or nonadjustable parameters. If there are nondeterministic components (i.e., if $\boldsymbol{\Theta}_k$ is not of length zero), a predictive distribution $p(\boldsymbol{\varphi}|M_k)$ will arise instead of deterministic predictions $\boldsymbol{\varphi}_k$.

We assume that the system state $\mathbf{y}_{true}$ is known from observations $\mathbf{y}_o$ up to a measurement error ϵ_o:

$$\mathbf{y}_o = \mathbf{y}_{true} + \epsilon_o. \tag{2}$$

Assumptions on the characteristics of the measurement noise can be formulated as $p(\epsilon)$. The model M_k will predict the system state as:

$$\mathbf{y}_k = M_k^{(y)}(\boldsymbol{\Theta}_k), \tag{3}$$

with $\mathbf{y}_k$ being the subset of $\boldsymbol{\varphi}_k$ that is used in model calibration. We distinguish between the prior predictive distribution $p(\boldsymbol{\varphi}|M_k)$ and the posterior predictive distribution $p(\boldsymbol{\varphi}|\mathbf{y}_o, M_k)$ after calibration with the observed data set $\mathbf{y}_o$. The Bayesian updating step to obtain posterior estimates based on a prior belief and the evidence in the calibration data is described in section 3.4. The expected value of the prior or posterior predictive distribution is denoted as $E[\boldsymbol{\varphi}|M_k]$ or $E[\boldsymbol{\varphi}|\boldsymbol{y}_o, M_k]$, respectively. The variance of these distributions is denoted by $V[\boldsymbol{\varphi}|M_k]$ (before calibration) and $V[\boldsymbol{\varphi}|\boldsymbol{y}_o, M_k]$ (after calibration).

2.2. Existing BMA Framework

The BMA equations are presented in detail in *Draper* [1995] and *Hoeting et al*. [1999], so that we will only briefly outline the existing BMA routine here. Note that model weights and weighted statistics are always conditional on the set of considered models. This is different from the interpretation of BME itself, as we will explain in section 2.3.

The model-averaged posterior predictive distribution of $\boldsymbol{\varphi}$ after calibration on $\mathbf{y}_o$ can be expressed as

$$p(\boldsymbol{\varphi}|\mathbf{y}_o) = \sum_{k=1}^{N_m} p(\boldsymbol{\varphi}|\boldsymbol{y}_o, M_k)\ P(M_k|\mathbf{y}_o), \tag{4}$$

with $P(M_k|\mathbf{y}_o)$ being posterior model weights.

The model-averaged posterior mean of $\boldsymbol{\varphi}$ is obtained by

$$E[\boldsymbol{\varphi}|\mathbf{y}_o] = \sum_{k=1}^{N_m} E[\boldsymbol{\varphi}|y_o, M_k]P(M_k|y_o), \tag{5}$$

and its posterior variance by

$$V[\boldsymbol{\varphi}|\mathbf{y}_o] = \sum_{k=1}^{N_m} V[\boldsymbol{\varphi}|\mathbf{y}_o, M_k]P(M_k|\mathbf{y}_o) + \sum_{k=1}^{N_m} (E[\boldsymbol{\varphi}|\mathbf{y}_o, M_k] - E[\boldsymbol{\varphi}|\mathbf{y}_o])^2 P(M_k|\mathbf{y}_o), \tag{6}$$

with the first term representing within-model variance (due to the uncertainty encoded in the probability distributions of $\boldsymbol{\Theta}_k$) and the second term representing between-model variance (conceptual uncertainty within the set $\mathcal{M}$ of considered models). Both terms result from applying the law of total variance with respect to the conceptual uncertainty within $\mathcal{M}$.

All the above equations require knowledge of the model weights (posterior probability of each model M_k), which are given by Bayes' theorem as

$$P(M_k|\mathbf{y}_o) = \frac{p(\mathbf{y}_o|M_k)P(M_k)}{\sum_{i=1}^{N_m} p(\mathbf{y}_o|M_i)P(M_i)}. \quad (7)$$

$P(M_k)$ is the prior probability (prior belief) that model M_k is the most adequate one in the set before considering the calibration data set $\mathbf{y}_o$. The BME term $p(\mathbf{y}_o|M_k)$ as introduced in section 1 quantifies the marginal (average) likelihood of the observed data given the model's parameter and input space [e.g., *Kass and Raftery*, 1995]:

$$p(\mathbf{y}_o|M_k) = \int_{\Omega_k} p(\mathbf{y}_o|M_k, \mathbf{\Theta}_k)p(\mathbf{\Theta}|M_k)d\mathbf{\Theta}_k. \quad (8)$$

$p(\mathbf{\Theta}|M_k)$ denotes the prior distribution of the model inputs and parameters, defined on the domain Ω_k. $p(\mathbf{y}_o|M_k, \mathbf{\Theta}_k)$ is the likelihood of a realization of model M_k to have generated the observed data set $\mathbf{y}_o$. For a comparison of available techniques to evaluate the BME term, be referred to *Schöniger et al.* [2014]. Here we perform a brute-force MC integration over each model's parameter and input space to obtain the BME values. Finally, the denominator in equation (7) normalizes the model weights such that they sum up to one.

2.3. Interpretation of BME and Posterior Model Weights

BME is an objective measure of the total likelihood that a specific model has generated the observed data and is not conditional on any set of competing models under consideration. This measure allows an objective comparison of model performance for a given data set and a specific definition of likelihood (i.e., for a specific assumption of measurement error statistics). BME is "future-proof" [*Skilling*, 2006] in that future models can be compared to the current one (by using the same data set $\mathbf{y}_o$ and likelihood formulation) without having to recalculate BME for the current model(s) under consideration. This does not hold for posterior model weights, as they are conditional on the set of models under consideration due to their joint normalization to sum up to one.

The ratio of BME for two competing models is equal to the Bayes factor [*Kass and Raftery*, 1995]. The Bayes factor $BF(M_k, M_l)$ is defined as ratio between the posterior and prior odds of model M_k being the more adequate one in comparison to model M_l:

$$BF(M_k, M_l) = \frac{P(M_k|\mathbf{y}_o)}{P(M_l|\mathbf{y}_o)} \frac{P(M_l)}{P(M_k)} = \frac{p(\mathbf{y}_o|M_k)}{p(\mathbf{y}_o|M_l)}. \quad (9)$$

The Bayes factor is a measure for significance in Bayesian hypothesis testing. It quantifies the evidence (literally, as in Bayesian model evidence) of hypothesis M_k against the null-hypothesis M_l.

Jeffreys [1961] provided a rule of thumb for the interpretation of Bayes factor values on a $\log_{10}$-scale. Later, *Raftery* [1995] slightly modified these grades of evidence with respect to $2\ln BF(M_k, M_l)$ for easier comparison with approximate BME values obtained from information criteria (see section 1). Following the interpretation of *Jeffreys* [1961], a Bayes factor of 1–3 represents evidence in favor of M_k that is "not worth more than a bare mention," a factor of up to 10 represents "substantial" evidence, a factor between 10 and 100 resembles "strong" evidence and, finally, a Bayes factor greater than 100 can be considered "decisive" evidence, i.e., it can be used as a threshold to reject models based on poor performance in comparison to the best performing model in the set. These thresholds should be seen as suggestions that can be of help in any specific application. Ultimately, it is the modeler's decision to define a confidence level at which they trust to reject a model or to choose a single best one out of the set. We will use the thresholds suggested by *Jeffreys* [1961] for illustration in our application (section 3).

2.4. Extension of BMA Framework to Account for Weighting Uncertainty

Weighting uncertainty exists if the model weights depend on any uncertain quantity ω. If both the calculated BME value and the derived model weights change with random outcomes of the uncertain quantity, they have to be viewed as random variable functions of ω.

Since BME is the likelihood of a model to have generated the observed data set, averaged over its parameter and input space (equation (8)), it is a function of the observed data values. The observed data values $\mathbf{y}_o$, in turn, are conceptualized as the sum of true system states and a realization of random measurement error

(equation (2)). BME thus depends on the random outcome of measurement noise and becomes a random variable function of $\omega=\epsilon$.

Model weights could also become random variables due to sources of uncertainty captured in the definition of $p(\mathbf{\Theta}|M_k)$. Instead of integrating over all relevant types of uncertainty (e.g., parameter uncertainty, input uncertainty, etc.) to obtain a single BME value per model (equation (8)), one could choose to investigate the variability in model weights due to any specific source of uncertainty by determining a BME value for different outcomes of the corresponding uncertain variable.

As an example, we might wish to assess the variability in model weights due to uncertain inputs **v**. In that case, we would remove **v** from $\mathbf{\Theta}_k$ when evaluating the BME integral and determine BME as a random variable function of $\omega=\mathbf{v}$. In the following, $\mathbf{\Theta}_k$ will refer to only those components which are marginalized over to obtain a BME value, while ω will refer to the uncertain variable whose impact on model weights shall be investigated.

We propose to quantify weighting uncertainty by running a resampling analysis within a Monte Carlo framework. We repeatedly draw from the assumed distribution $p(\omega)$ to generate an ensemble of N_d realizations. As a general statistical analysis tool, this type of resampling strategy is referred to as parametric bootstrap [e.g., *Davison and Hinkley*, 1997]. For each random realization of ω drawn from $p(\omega)$, the corresponding BME value is determined by integration over the model's (remaining) parameter and input space $\mathbf{\Theta}_k$ (equation (8)). Model weights are then determined from equation (7). Repeating this analysis for the different realizations of ω drawn from $p(\omega)$ yields distributions of BME values and model weights. Based on these distributions, we assess the (non)robustness of the BMA ranking against random outcomes of the uncertain variable ω.

2.5. Theoretical Upper Limit for Model Performance due to Measurement Noise in Output Data

In the specific case of measurement noise in output data (i.e., in the calibration data set), a theoretical limit for model performance exists. We propose to identify this limit by determining a distribution of BME for a *theoretically optimal model* (TOM). We define the observed data set as TOM, since it shows a perfect fit (zero bias) while using a minimum number of parameters (exactly zero, equivalent to zero variance). *Jaynes* [2003] calls this the *sure thing* hypothesis. Refer to Appendix A for an explanation how the sure thing hypothesis would perform in a BMA analysis, if it was actually formulated as an alternative model. Here we do not include the TOM into the actual model ranking, but use it only as an upper limit to the BME scale: the TOM represents the best possible performance in presence of measurement error.

We determine the theoretically optimal distribution of BME under measurement noise as the distribution of likelihoods of the observed data set given the perturbed data sets (details on the numerical implementation are provided in section 3.5). Instead of again using the parametric bootstrap approach to obtain this distribution, we could also make use of an analytical expression for the case of independent and identically distributed Gaussian measurement errors: the TOM performance (expressed as log-BME) is then a distribution of the weighted sum of normal squared residuals, and is therefore defined by the chi-square distribution [e.g., *Hald*, 1998] with the number of degrees of freedom being equal to the size of the calibration data set N_s. Hence, the upper limit of performance as represented by the TOM is imposed by the presence of measurement noise, but the limit does *not* depend on the actual level of measurement error variance, because the chi-square distribution is only a function of the data set size.

2.6. Measures to Compare Resulting BME Distributions

The BME distributions of the competing models can be compared with each other or (in the special case of addressing measurement noise in output data) with the optimal distribution of BME (obtained for the TOM) either by visual inspection of probability density graphs, or by quantitative measures. Since likelihoods tend to span a very large range of values and typically show a skewed distribution with the largest mass close to zero, we analyze the differences in distributions of log-BME instead. The interpretation of BME on a log-scale is also more intuitive, because on that scale it is comparable to the sum of squared errors that would be typically evaluated during calibration (see definition of likelihood function in equation (14)). Further, approximations to the true BME value by information criteria act on this scale (see section 1). Probability density functions of log-BME, $p(Y)$ with $Y=\log_{10}$ (BME), can be obtained from the ensemble of N_d BME values per model by kernel density estimation [e.g., *Bowman and Azzalini*, 1997].

The difference between probability distributions can be quantified with various distance measures which differ in their properties and interpretation. We discuss three alternatives in the following. As a quantitative measure for the distance between distributions of log-BME, one could determine the distance D_{mode} between the modes $\tilde{Y}$ of the respective log-BME densities, with $\tilde{Y} = \max_Y p(Y)$:

$$D_{mode}(M_k, M_l) = \tilde{Y}_{M_k} - \tilde{Y}_{M_l}. \tag{10}$$

As an alternative, the dissimilarity of the two probability density functions can be quantified with the Kullback-Leibler (KL) divergence [*Kullback and Leibler*, 1951] D_{KL}:

$$D_{KL}(M_k, M_l) = \int p(Y|M_k) \ \ln\left(\frac{p(Y|M_k)}{p(Y|M_l)}\right) dY. \tag{11}$$

The KL divergence is frequently applied in the context of information theory as a measure for the information lost by approximating one probability density function with a different one (see, e.g., *Nearing et al*. [2013] for an application to crop modeling).

Similar to the difficulties in interpreting BME values (section 2.3), these distance measures do not have an intrinsic meaningful scale to them. A dimensionless measure with fixed bounds between zero (minimum distance) and one (maximum distance) might provide a more intuitive basis to judge model performance. The Hellinger distance [*Hellinger*, 1909] is such a relative measure. It is defined as

$$D_H(M_k, M_l) = \sqrt{1 - BC(p(Y|M_k), p(Y|M_l))}, \tag{12}$$

with BC representing the Bhattacharyya coefficient [*Bhattacharyya*, 1943]:

$$BC(p(Y|M_k), p(Y|M_l)) = \int \sqrt{p(Y|M_k) p(Y|M_l)} dY. \tag{13}$$

The Hellinger distance enables a modeler to compare arbitrary models (including the TOM) with respect to arbitrary data sets, with the minimum distance (i.e., for completely identical distributions) being zero, and the maximum distance (i.e., no overlap at all) being one. Using the Hellinger distance to assess the distance between a model and the TOM is similar to the approach of *Abramowitz and Gupta* [2008], who define a slightly different distance measure based on the overlap of the predictive distributions of competing models. While their primary intention is to assess the independence of competing models (i.e., their dissimilarity), they also mention that one could use such a measure to assess the model performance against the observed data set.

2.7. Comparison of Resulting BME Values Through Bayes Factors

Independently from the chosen measure to quantify the overlap between BME distributions, this overlap itself might be misleading: we do not know whether the overlap actually occurs for individual realizations of the uncertain variable ω, or whether the overlap is artificial in the sense that it reflects BME values based on different samples (and thus we would never actually see this overlap in model ranking no matter what the outcome of the uncertain variable was). To distinguish between those cases, one can again use the Bayes factor (equation (9)) and apply it to the BME values corresponding to an individual realization j from the ensemble of $j = 1 \ldots N_d$ randomly drawn samples of the uncertain variable.

Note that this procedure directly applies to BME values instead of their log-transform. The resulting distribution of Bayes factors for the ensemble of samples then also spans a wide range of values, so that we again report them in $\log_{10}$-space. This does not impact the comparison of performance, but only helps in visualization, as opposed to taking the logarithm of BME before estimating probability density functions. The closer the log-BF distribution of a model stays to a Bayes factor value of one (or log-BF of zero), the closer it is to the compared model. In other words, two models are not distinguishable when they obtain similar BME values.

From a hypothesis testing perspective, a modeler could ask at which significance level they should reject a model against a competing model or against the TOM. This significance level can be determined from the cumulative distribution function of log-BF, using the cumulative density value that corresponds to a Bayes factor of, e.g., 100 (decisive evidence) [*Jeffreys*, 1961]. Such a significance test is our preferred tool to discover performance differences between models and to monitor improvements in model building. The

suggested way to report such findings is that "on an x% significance level, we can/cannot conclude from the available data that there is decisive Bayesian evidence against model Y when compared to model Z."

3. Application to Soil-Plant Model Selection

In this section, we apply the proposed extended BMA routine to soil-plant model selection. Our goal is to assess the plausibility of four competing crop models which predict different aspects of the soil water balance and of plant growth. We first give a brief overview of soil-plant modeling in section 3.1. We then describe the field site, the experimental setup, and the considered models in section 3.2. In section 3.3, we provide details about the data used to calibrate the models. The numerical implementation of the existing BMA framework is outlined in section 3.4, and the additional implementation steps of our suggested extension are explained in section 3.5. Results from the BMA analysis, first without and then with our suggested extension, are presented in section 4.

3.1. Approaches to Soil-Plant Modeling

Soil-plant interactions are a field of intensive research [see, e.g., *Molz*, 1981] because they govern exchange fluxes of water, nutrients and energy at the land surface. To adequately capture those feedback processes between the land surface and the atmospheric boundary, integrative approaches are needed that represent the soil-plant-atmosphere system as a continuum. Various different modeling approaches have been applied to address this challenge [e.g., *Bonan et al*., 2002; *Priesack and Gayler*, 2009; *Niu et al*., 2011; *Greve et al*., 2013]. These modeling approaches vary significantly in how detailed vegetation and soil processes are resolved. Typically, land surface schemes that are coupled to atmospheric models for climate studies consider surface vegetation processes in detail, but lack a detailed representation of subsurface hydrological processes and soil-plant interactions. Agricultural models (crop models) for simulating water and nitrogen budgets, on the other hand, act on a much smaller scale and typically represent soil processes and root water uptake in more detail [*Gayler et al*., 2013].

As expected from very differently elaborate modeling approaches, the prediction accuracy of soil-plant models varies substantially. In particular, differences in representing the dynamic plant processes lead to large differences in prediction quality [*Wöhling et al*., 2013]. In a study using a model ensemble of the Noah-MP multiphysics land-surface scheme, *Gayler et al*. [2014] reported large variability in the performance of the individual models to simulate water and energy fluxes. *Priesack et al*. [2006] showed that the choice between competing crop models has a strong influence on predictions of carbon and nitrogen turnover. This conceptual choice between competing models also influences predictions of environmental impacts on crops [*Biernath et al*., 2011].

Various studies have revealed that competing soil-plant models possess different structural deficiencies which depend on the actual field site conditions [*Nearing et al*., 2012; *Wöhling et al*., 2013; *Houska et al*., 2014]. Therefore, model choice cannot be confidently done prior to the application, but must be performed based on an actual calibration data set from the specific site of interest. This calls for an approach such as BMA. *Wöhling et al*. [2015] have presented a first application of BMA to soil-plant modeling which provided many valuable insights into model structural deficiencies of four selected crop models. *Wöhling et al*. [2015] further investigated the impact of different data types (evapotranspiration rate, soil moisture content, and leaf area index) and different subsets and combinations thereof on the outcome of model weights. They found that model ranking can vary significantly with varied calibration data sets.

We continue these investigations in this study with a focus on the impact of measurement noise in the calibration data set. Our goal is to assess the variability in model weights due to measurement noise, and the resulting consequences on model ranking.

3.2. Description of Soil-Plant Models and Field Site

In this study, the same four alternative models as in *Wöhling et al*. [2015] are used to predict latent heat flux (actual evapotranspiration, ET_a), leaf area index (LAI, the leaf area per square meter ground surface), soil moisture, and drainage at a field site at the Swabian Alb (48.510°N and 9.810°E, 690 m a.s.l.) in South-West Germany. On this field, wheat is grown on a shallow and stony Rendzina soil with a solum thickness of 0.2–0.3 m. Soil management practice includes fertilization and crop protection as usual in conventional farming in this region. The complete information about the site properties can be found in *Wizemann et al*. [2015].

The vegetation components of the four different crop models are combined with identical routines for simulating soil water movement [*Simunek et al*., 1998], soil hydraulic properties [*Van Genuchten*, 1980], soil carbon and nitrogen turnover [*Johnsson et al*., 1987], and soil heat and nitrogen transport [*Hutson and Wagenet*, 1992] to simulate vertical transport of water, solute and heat in the unsaturated zone, organic matter turnover, and crop growth. The four crop models are CERES-WHEAT (subsequently abbreviated CERES) [*Ritchie et al*., 1988], SPASS [*Wang and Engel*, 2000; *Gayler et al*., 2002], SUCROS2 (subsequently abbreviated SUCROS) [*van Laar et al*., 1997], and GECROS [*Yin and van Laar*, 2005] and are used as implemented in the modular model system Expert-N 3.0 [*Priesack et al*., 2006].

Distinct differences exist between the four models in simulating root water uptake and root development, leaf-area growth, photosynthesis, and stomatal resistance. Since these four models have been used in several simulation studies before [*Wöhling et al*., 2013; *Gayler et al*., 2014; *Wöhling et al*., 2015], the reader is referred to these studies for detailed information on the individual model structures. In brief, SPASS and CERES use similar routines for root growth and root water uptake, while SUCROS and GECROS do not allow as much flexibility in the vertical root distribution. The leaf area growth, photosynthesis, and plant internal distributions of assimilates and nitrogen are simulated in most detail by GECROS, while SPASS is a hybrid model composed of parts from CERES and SUCROS.

Consistent with our earlier modeling approach [*Wöhling et al*., 2015], we assume two horizontal soil layers in all our simulations. As uncertain parameters in that modeling scheme, we use the same ones as identified in the cited previous works. These are five common soil hydraulic parameters for each of the two horizons and three to four crop model parameters, specific to each model. Specifically, the hydraulic parameters to be calibrated are the saturated water content θ_s (m^3/m^3), the van Genuchten shape parameters of the water retention function α and n, the saturated hydraulic conductivity K_s (cm/d), and the pore-connectivity parameter l [*Van Genuchten*, 1980]. The crop parameters for the CERES and SPASS models are the maximum root extension rate δ_r (cm/d), the specific root length density λ_R (m/kg), the maximum water uptake rate per root length ξ_W (cm^3/cm/d), and the specific leaf weight λ_L (kg/(ha leaf area)). In SUCROS, ξ_W is not considered, because the model does not use a parameter for the maximum water uptake rate. The GECROS crop parameters used for calibration are the specific leaf area s_{la} (m^2/(g leaf)), the critical root weight density w_{Rb} (g/m^2/(cm depth)), the minimal leaf nitrogen $n_b = 0.01\varepsilon/s_{la}$ ((g nitrate)/m^2), and the slope of the maximum carboxylation rate versus leaf nitrogen $\Delta_{Vc,max}$ (μmol/s/(g nitrate)).

The uncertainty in input data is comparably small for this specific application because the meteorological data were measured on-site. Hence, we treat model inputs $\mathbf{v}$ as deterministic (i.e., they belong to the set of deterministic variables $\mathbf{c}$, cf. section 2.1). Further, there is no stochastic component $\mathbf{w}$ in the modeled system. For the soil-plant models considered here, structural error models have not yet been developed and will hence not be considered. This leaves us with $\boldsymbol{\Theta}_k = \mathbf{u}_k$ (cf. section 2.1), and model predictions are obtained as predictive distributions due to the uncertainty in parameters $\mathbf{u}_k$.

3.3. Data Used for Calibration of the Four Soil-Plant Models

We use the same data to calibrate the four models as in *Wöhling et al*. [2015], obtained from field experiments by *Wizemann et al*. [2015] during a growing season of winter wheat (October 2008 to August 2009). The site was equipped with an eddy-covariance tower to measure energy and water fluxes between canopy and atmosphere. ET_a data were gap-filled as described by *Wizemann et al*. [2015] and aggregated to weekly averages of daily values for use in model calibration. Soil moisture measurements were taken at two different depths (i.e., 5 and 15 cm) using TDR probes and aggregated to daily values. LAI was measured at five subplots in biweekly intervals until grain maturity. LAI averages from the five subplots were used for model calibration.

The four soil-plant models are simultaneously calibrated on different data types, i.e., on either eight observations of LAI (scenario 1), on 16 weekly averages of daily actual ET_a rates (scenario 2), or on both data types at the same time (scenario 3). More details on calibration strategies of these four competing models can be found in *Wöhling et al*. [2013, 2015]. Based on these different calibration scenarios, we investigate to which extent the impact of measurement noise on model ranking depends on the calibration variable.

The measurement error standard deviation for each LAI observation was determined from replicated field measurements. For ET_a, 15% of the average measured value is assumed as measurement error standard

deviation to cover the combined effect of inaccuracies in eddy covariance measurements and deficiencies in reconstruction and aggregation of diurnal variations.

For the time period analyzed in this study, there were no drainage observations available. The four models could also be calibrated on soil water content, which is of great value to constrain the uncertainty in the water balance, but was shown to be of lesser importance for the purpose of model selection than LAI and ET_a [*Wöhling et al.*, 2015]. This finding is related to the fact that all four models share the Richards' equation to simulate water movement through the soil. In this study, we focus on model selection in light of LAI and ET_a data and do not consider observations of soil water content.

3.4. Numerical Implementation of Existing BMA Routine

To set up the Monte Carlo framework for calibration and for model ranking, we have generated an ensemble of predictions based on N_{MC} random realizations of parameters for each of the four soil-plant models. Uncertain parameters were drawn randomly from uniform priors $p(\mathbf{u}_k|M_k)$. The prior bounds were chosen to coincide with mostly physically motivated parameter limits [*Wöhling et al.*, 2013]. From a convergence analysis not shown here, we determined the final ensemble sizes which varied between 90,000 and 250,000 realizations per model.

We define the likelihood function $p(\mathbf{y}_o|M_k,\boldsymbol{\Theta}_k)$ in equation (8) (representing the distribution of measurement error) as normal distribution with covariance matrix **R**:

$$p(\mathbf{y}_o|M_k,\mathbf{u}_k)=2\pi^{-N_s/2}|\mathbf{R}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{y}_o-\mathbf{y}_k)^T\mathbf{R}^{-1}(\mathbf{y}_o-\mathbf{y}_k)\right]. \tag{14}$$

R is a diagonal matrix of size N_s (length of the calibration data set) $\times N_s$, with measurement error variances that are specific to each data type as described in the previous section.

To obtain posterior (calibrated) predictive distributions $p(\boldsymbol{\varphi}|\mathbf{y}_o,M_k)$, the prior realizations $\mathbf{u}_{k,i}$, with $i=1\ldots N_{MC}$, are weighted with their respective likelihood $p(\mathbf{y}_o|M_k,\mathbf{u}_{k,i})$ to have generated the calibration data set $\mathbf{y}_o$. This Bayesian updating procedure to calibrate models is referred to as weighted bootstrap [*Smith and Gelfand*, 1992].

BME is determined as the ensemble average of likelihoods (equation (8)), and posterior model weights are then obtained by normalization (equation (7)), assuming uniform (equal) prior model weights. The choice of neutral prior model weights is based on the verdict of a previous study [*Wöhling et al.*, 2013] that none of the four competing models is clearly to be favored or rejected as adequate model to simulate soil-plant-atmosphere interactions at this specific field site, before testing against site-specific calibration data.

3.5. Numerical Implementation of BMA Routine Extension

For this specific application, we are interested in the impact of measurement noise in the calibration data set on the outcome of model weights and of model ranking. We thus address the issue of weighting uncertainty caused by noise in the calibration data set and implement the extended BMA routine as proposed in section 2.4 to account for the uncertainty in $\omega=\epsilon$. Treating this source of uncertainty first is also a natural first step in extending the existing BMA routine, because assumptions on the error in the observations of model output variables already enter the existing BMA analysis through the definition of the likelihood function. For other case studies or other systems (e.g., rainfall-runoff models), input uncertainty might be of higher importance and could be treated with our suggested extension as well (cf. section 2.4).

We extend the existing BMA routine to explicitly account for measurement noise by perturbing the observed data with random realizations of measurement error in a parametric bootstrap approach. We use exactly the same assumptions for measurement error as in the definition of the likelihood function (equation (14)). We now add random realizations of measurement error according to the specified distribution:

$$\mathbf{y}'_o=\mathbf{y}_o+\boldsymbol{\epsilon}, \tag{15}$$

with $\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{R})$.

If we knew the exact outcome of measurement error for the original data set, we would have to subtract that noise before adding the random noise (but then again, this would make the whole weighting uncertainty analysis dispensable). In absence of further knowledge, however, the assumption of (on average) zero

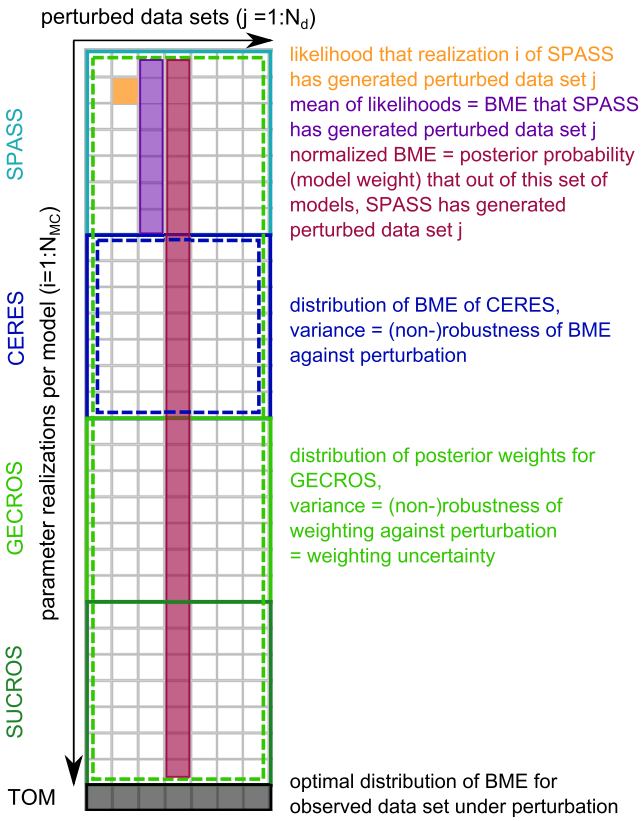


Figure 1. Schematic illustration of resampling analysis within a numerical Monte Carlo framework for BMA.

noise is the statistically correct starting point. In statistical terms, we use the error-prone data values as point estimate for the true system states.

Figure 1 schematically shows how the extended BMA routine is implemented for the four models considered here. The columns of the matrix represent different realizations of measurement error that are added to the original calibration data set. The rows represent different realizations of parameter values for each of the four models. Each cell with index *i*, *j* contains the likelihood that the specific parameter realization $\mathbf{u}_{k,i}$ of model M_k has generated the specific perturbed data set $\mathbf{y}'_{o,j}$. Averaging the likelihoods $p\left(\mathbf{y}'_{o,j}|M_k,\mathbf{u}_{k,i}\right)$ of all parameter realizations *i* of model M_k yields its BME value, i.e., the integrated likelihood that this model has generated the specific data set $\mathbf{y}'_{o,j}$. Normalizing the *k* BME values with their sum for the same perturbed data set yields posterior model weights (equation (7)). Repeating this analysis for all perturbed data sets yields distributions of BME values and model weights. Based on these distributions, we assess the (non)robustness of the BMA ranking against random outcomes of measurement error.

Repeating the BMA analysis for a number of N_d perturbed data sets results in $N_{MC}\times N_d$ evaluations of likelihood per model, with N_{MC} being the prior ensemble size. Typically, the number of perturbations N_d required for convergence is much smaller than the prior ensemble size (see, e.g., remarks by *Leube et al*. [2012] in the context of Preposterior Data Impact Assessment). Also, the BMA analysis routine which needs to be repeated N_d times is typically much less computationally demanding than the model runs to generate predictions from the prior ensemble. Therefore, our suggested extension of the BMA routine only requires a small additional amount of computational power if the BMA analysis has already been performed in a Monte Carlo framework. The latter is highly recommended to obtain accurate model ranking results [*Schöniger et al*., 2014].

We generate 10,000 perturbed data sets of LAI (calibration scenario 1) and ET_a (calibration scenario 2), and 50,000 data sets that consist of perturbed LAI and ET_a data (calibration scenario 3). Note that we do not assume any correlation between measurement errors of LAI and ET_a in scenario 3. We have determined the required number of samples from a convergence analysis not shown here. We choose relatively large ensembles of perturbed data sets here to ensure stable statistics of BME. In practice, smaller ensembles would still give an indication to which extent BMA weights suffer from noise in the observation data for the application at hand.

For each perturbed data set, we repeat the BMA analysis and store the resulting BME values to determine distributions of Bayes factors and distance measures between the models and the TOM (the original unperturbed data set) as proposed in sections 2.6 and 2.7. We obtain all probability density functions from kernel density estimation based on the ensembles of log-BME or log-BF, using Gaussian kernels with an optimal kernel width [*Bowman and Azzalini*, 1997].

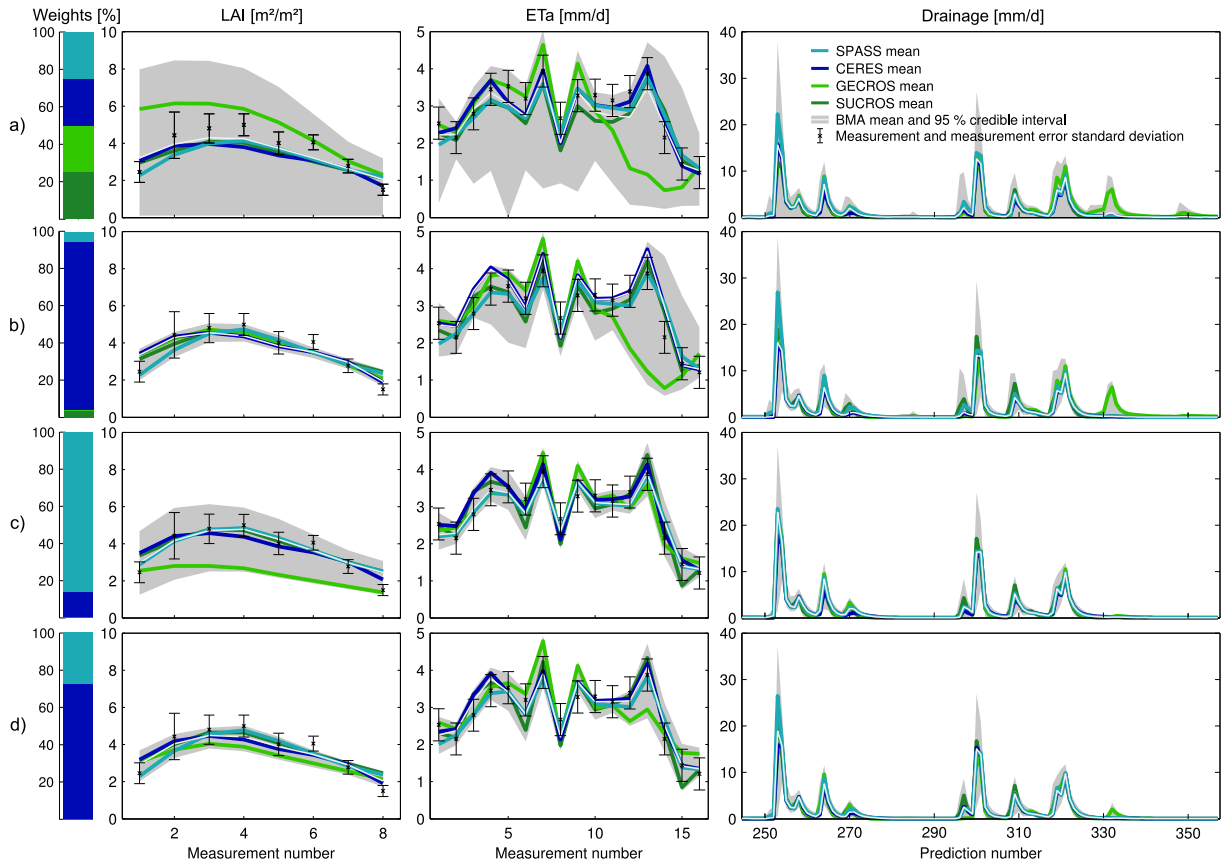


Figure 2. Model weights, model predictions, and BMA-weighted predictions of LAI, ET_a, and drainage. (a) Prior state of knowledge, (b) posterior state after calibration on LAI data, (c) posterior state after calibration on ET_a data, and (d) posterior state after calibration on both LAI and ET_a data.

It is important to note that explicitly accounting for measurement noise as source of uncertainty for model weights is not "double accounting" for measurement noise. Assumptions on the distribution of measurement error already enter the BMA procedure through the definition of the likelihood function, which is required to determine BME from equation (8). The likelihood function acknowledges that a model prediction which deviates from the observation in a plausible range (defined by the assumed distribution of noise) is not inherently false, but could be correct with a corresponding probability. However, the whole BMA analysis builds on the actually observed data set, so that matching the observed data will produce the highest likelihood. Our approach now takes a frequentist perspective on this by asking the question: What would be the outcome of the model weights, if it was based on a set of repeated measurements with a new random outcome of measurement error, i.e., if the maximum likelihood prediction was shifted?

4. Results and Discussion

4.1. Individual Model Performance Before and After Calibration

The prior (uncalibrated) and resulting posterior (calibrated) mean predictions for LAI, ET_a, and drainage as obtained from the four competing crop models are shown in Figure 2. These model means represent the expected value over the model's prior or posterior predictive distributions, respectively (see section 3.4) Note that we obtain predictive distributions instead of deterministic predictions because we consider parameter uncertainty within the models.

We provide measurement numbers instead of dates, but measurements span the same time period and are approximately equally spaced in time, so that their interactions become obvious. For example, the low ET_a predicted by GECROS before calibration (Figure 2a) toward the end of the season (measurement number

14) results in a peak of drainage (measurement number 332) which is only predicted by GECROS. The other models predict a higher ET_a and therefore a higher loss of water to the atmosphere. Observations and their assumed standard errors (±one standard deviation) are also marked in Figure 2.

All of the models follow the general trend of increasing LAI and ET_a values in spring, with highest values at the peak of the cropping season and decreasing values at senescence, and some weather-related short-term variations which are more pronounced in ET_a rates than in LAI. Thus, all four models are in general able to represent the dynamics of the featured soil-plant-atmosphere system well. However, differences in model performance are obvious from the prior predictive distributions at specific data points in time, e.g., GECROS neither reproduces the ET_a peak in measurements number 11 and 12 nor the low LAI values at the beginning of the season. These differences partly diminish during calibration, depending on the chosen calibration scenario.

4.2. Model Ranking Based on Existing BMA Routine

The resulting model weights based on the different calibration scenarios as defined in section 3.3 are shown on the very left side of Figure 2. These weights were determined and discussed in the study of *Wöhling et al*. [2015]. In general, *Wöhling et al*. [2015] found that the BMA-weighted prediction achieved a better fit to the observed data than the individual models.

The specific impact of the model weights per calibration scenario on the respective BMA mean and the spread of the BMA-weighted model ensemble (as measured by 95% credible intervals of the model-averaged predictions) can be seen in Figure 2. For example, one can observe the influence of GECROS on the overall BMA prediction: GECROS obtains a nonnegligible weight only when calibrated on LAI alone (Figure 2b). Only then, one can find the peak in drainage (measurement number 332) again within the posterior credible intervals of the BMA-weighted multimodel ensemble, which is shaded in gray in Figure 2. Despite its very small model weight, GECROS is obviously able to significantly influence the model-averaged predictions here. However, credible intervals provide a measure for the spread of the multimodel ensemble; a closer look at the probability density of the prediction would reveal that only a very small fraction of the BMA ensemble actually predicts such a spike in drainage.

Wöhling et al. [2015] identify two reasons why GECROS obtains a very small or even negligible posterior model weight in the calibration scenarios considered here: First, GECROS shows a strong trade-off in the simultaneous fit to ET_a and LAI data, i.e., it fails in accurately predicting ET_a when calibrated on LAI only, and vice versa. Second, it shows a larger prediction bias in all data types than the other three models. The authors hypothesize that the worse performance might be related to the shallow soil under consideration, because, for a different field site with a deep loess soil, GECROS performed better [*Wöhling et al*., 2013]. These findings indicate that the assessment of competing crop models should be repeated for different sites and conditions before making a final judgment about the usefulness of each model for future applications.

The reduction in uncertainty (the shrinkage of credible intervals of the model-averaged predictions) clearly depends on the calibration scenario: it is highest if the respective variable has been used for calibration, and lowest if not. Prediction uncertainty is similarly small when calibrating on both LAI and ET_a simultaneously. The predictive uncertainty in drainage can be significantly reduced by calibration on either LAI, or ET_a, or both. All three calibration scenarios lead to much more pronounced dynamics in the posterior predictions. The peak flows, however, still show large credible intervals, which is due to the fact that the water balance has not been further constrained by soil moisture measurements in the calibration scenarios considered here. Refer to *Wöhling et al*. [2015] for a discussion of results when using soil moisture measurements for calibration.

4.3. Variability in Model Weights due to Measurement Noise in Output Data

From the resampling analysis described in sections 2.4 and 3.5, we obtain a set of 10,000 model weight combinations in the case of the calibration scenarios 1 (only perturbed LAI data) and 2 (only perturbed ET_a data) and 50,000 combinations in the case of calibration scenario 3 (perturbed LAI and perturbed ET_a data).

Figure 3 illustrates the resulting variability in posterior model weights. Figure 3a shows 50 arbitrarily picked outcomes of model weights based on calibration scenario 1 that could result from using this experimental setup (i.e., the chosen measurement locations and times to collect LAI data) to perform BMA.

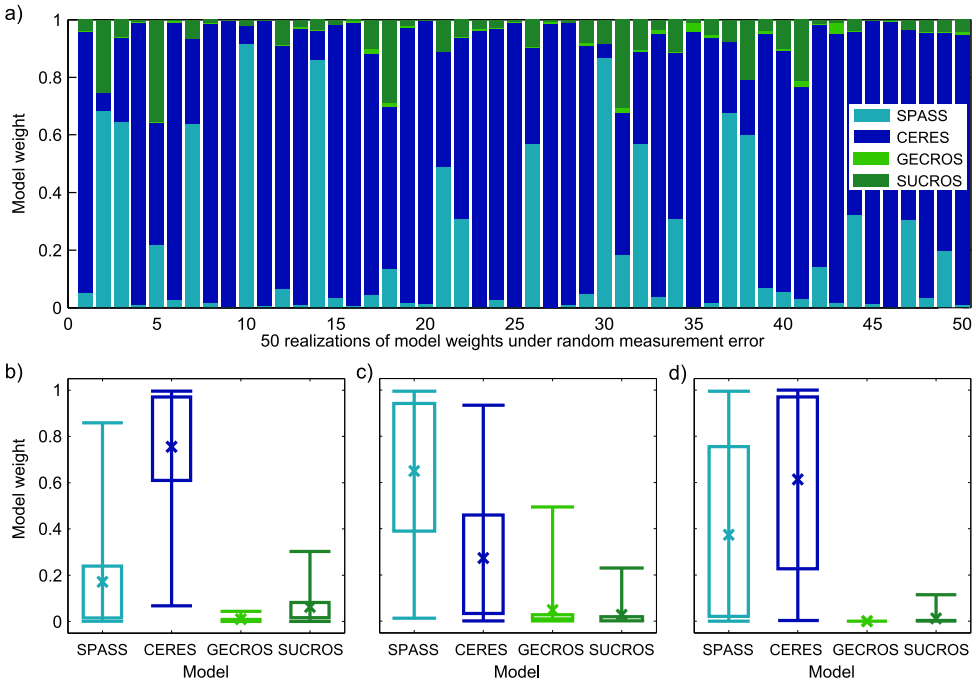


Figure 3. Variability in posterior model weights resulting from resampling analysis. (a) Arbitrarily chosen sets of model weight outcomes during perturbation of LAI data (calibration scenario 1). Mean model weights (crosses), 50% confidence intervals (boxes), and 95% confidence intervals (whiskers) after (b) 10,000 perturbations based on calibration scenario 1, (c) 10,000 perturbations based on calibration scenario 2 (using ET_a data), and (d) 50,000 perturbations based on calibration scenario 3 (using both LAI and ET_a data).

Obviously, there is a large variability in model weights and model ranking, which is summarized statistically in Figure 3b. While the mean model weights under the impact of measurement noise convey a relatively clear model ranking in favor of CERES, with SPASS ranking second and GECROS clearly ranking last, this model ranking does not prove to be reliable: given different outcomes of measurement error, the frequentist confidence intervals for the model weights show that SPASS could also be ranking first, CERES could obtain a smaller model weight than SUCROS, and so forth. This means that under the given measurement noise assumptions, model ranking could also be turned upside-down. The only exception is GECROS, which shows a very small uncertainty in its very small weight. In this case, the modeler would concede that this experimental setup is not a reliable basis for model selection or model averaging. Note that the mean model weights determined from the resampling analysis (Figure 3b) do not coincide with the original weighting based on the unperturbed data set (Figure 2b) due to the extreme nonlinearity in the likelihood function.

Similar conclusions are drawn from the resampling analysis based on calibration scenarios 2 (using ET_a data) and 3 (using both LAI and ET_a data). The resulting variability in posterior model weights is shown in Figures 3c and 3d, respectively. The ambiguity in model ranking between SPASS and CERES is even larger than in the case of calibration on LAI data alone. In the case of calibration on ET_a data, SPASS ranks first in the majority of perturbations, while for calibration on both data types, there is a very large overlap even in 50% confidence intervals. Further, the 95% confidence interval of model weights for GECROS significantly widens when using ET_a data for calibration.

These differences in model ranking between different calibration scenarios are due to different model structural deficiencies that become evident in light of varied calibration data types as discussed by *Wöhling et al.* [2015]. Differences in the amount of weighting uncertainty can also be caused by the different assumptions on the distribution of measurement error. In the case of LAI data, measurement error statistics have been derived from replicated field measurements and are as such expected to be realistic (see section 3.3). For ET_a, measurement error statistics are chosen such that different potential sources of error are covered. These assumptions could be further scrutinized in order to obtain most realistic estimates of weighting

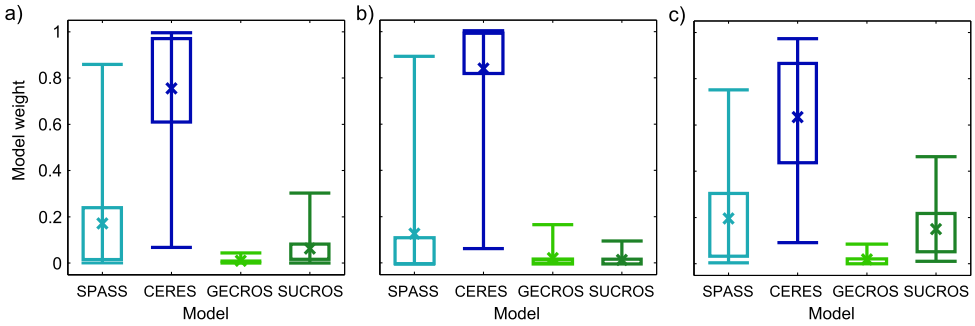


Figure 4. Variability in posterior model weights resulting from resampling analysis based on calibration scenario 1 (using LAI data) with (a) realistic measurement error variance, (b) reduced measurement error variance, and (c) increased measurement error variance.

uncertainty. Also, model structural errors could be introduced to clearly distinguish the mismatch between model predictions and ET_a data that is due to model bias from the mismatch attributed to random measurement errors.

To illustrate the influence of the assumed measurement error variance on model ranking statistics, we repeat the analysis for calibration scenario 1 (using LAI data) with decreased measurement error variance (half of the measurement error variance that was actually determined from replicated field measurements, see section 3.3), and with increased measurement error variance (twice the actually determined measurement error variance). The decrease or increase in measurement error variance affects both the generation of random realizations of measurement error for the resampling analysis, as well as the definition of the likelihood function (equation (14)) that is needed to determine BME (equation (8)). The covariance matrix $\mathbf{R}$ is now replaced by $\mathbf{R}^* = \frac{1}{2}\mathbf{R}$ in the case of reduced measurement noise and by $\mathbf{R}^* = 2\mathbf{R}$ in the case of increased measurement noise. Results are shown in Figures 4b and 4c, respectively, and can be compared to the base case of realistic measurement error variance in Figure 4a (equal to Figure 3b). As expected, model choice becomes more clear in the case of reduced measurement noise (or increased information in the data values). The mean model weights determined in the resampling analysis are more extreme, i.e., weights above 50% increase, weights below 50% decrease. The only exception is the very small mean model weight for GECROS which increases from 0.01 to 0.02. Also, the 50% confidence intervals of model weights shrink. In contrast, model choice becomes more ambiguous in the case of increased measurement noise (or less information in the data values), which can be seen from less extreme mean model weights and wider confidence intervals.

The impact of measurement noise on model ranking statistics further depends on the size and information content of the data set. In theory, the impact of measurement noise would average out over (infinitely many) repeated measurements or over a long time series of calibration data measured in a stationary system. Neither of these two conditions holds in our application, so that individual data points provide a variable information content for model ranking [see *Wöhling et al.*, 2015]. Since it would not be possible to separate the influence of increasing information content (in additional data points which sample a different state of the nonstationary system) from the decreasing impact of measurement noise (in additional data points which sample the same state and only differ in the outcome of measurement error), we do not investigate the influence of increasing time series length on model ranking under measurement noise in this study.

4.4. Confidence in Model Ranking in Presence of Measurement Noise in Output Data

The confidence in model selection can be analyzed with the help of Bayes factors (equation (9)) for pairwise competing models (i.e., SPASS against CERES, SPASS against GECROS, SPASS against SUCROS, CERES against GECROS, and so on). For each perturbed data set $\mathbf{y}'_{o,j}$, we obtain nine Bayes factors for each of the models against each of the other models. The probability density functions of log-BF over all N_d perturbed LAI data sets (calibration scenario 1) are plotted in an $N_m \times N_m$ matrix in Figure 5. This type of graph allows us to assess the pairwise ranking of each model compared to each of the other models. For example, we can

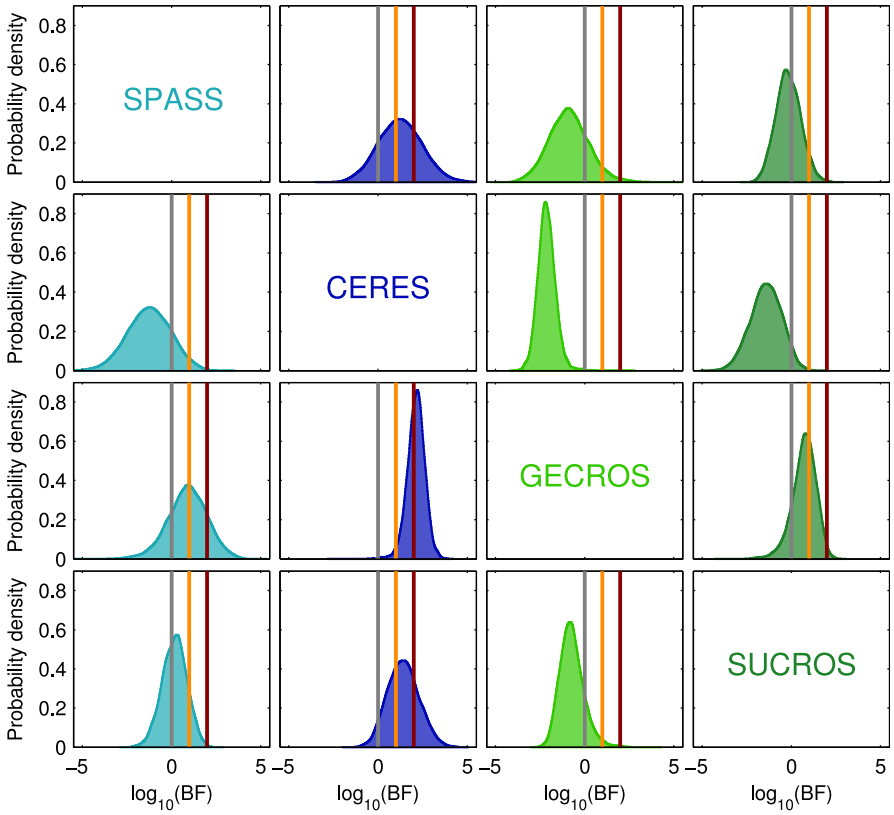


Figure 5. Distributions of $\log_{10}$(Bayes factor) for pairwise competing models based on calibration scenario 1 (using LAI data). The second plot in the first row, e.g., shows the distribution of $\log_{10}$(Bayes factor) in favor of CERES against SPASS. Gray lines indicate equally strong evidence for both models. Orange and red lines indicate thresholds for strong and decisive evidence (according to *Jeffreys* [1961]) in favor of one model against the other.

extract the information that CERES indeed achieves in almost all cases a higher BME value than SUCROS, because its Bayes factor (second column, fourth row in Figure 5) is greater than one (or log-BF greater than zero, indicated by the gray line) for most of the probability mass. The orange and red lines indicate significance levels as suggested by *Jeffreys* [1961] and discussed in section 2.3. Here in more than 50% of the analyzed cases (perturbed data sets), SUCROS could be rejected against CERES based on strong evidence. This might, however, not be convincing enough for a modeler to completely discard SUCROS from the model set. In contrast, a large fraction of Bayes factors of CERES against GECROS prove a decisive evidence (log-BF values greater than two in the second column, third row of Figure 5). Similarly, such a matrix of log-BF probability density functions could be constructed based on the calibration scenarios 2 and 3 (not shown here).

In general terms, this matrix allows the modeler to analyze the reliability in pairwise model ranking in great detail and thus provides a valuable tool for model selection, model building, and planning of experiments.

A decrease or increase in measurement noise tends to lead to stronger or weaker evidence in favor of the best model, respectively. This is in agreement with the findings presented in section 4.3. On the other hand, the distribution of Bayes factors strongly depends on the actual choice of competing models and of the data set (including its measurement errors). Any more general conclusions about the influence of measurement error variance on the confidence in model selection are therefore difficult to derive.

4.5. Comparison of Differences in Performance Between Models and TOM Based on BME Distributions

As explained in section 2.5, we can not only compare models against each other to decide if they should be discarded from the model set, but we can also compare their performance to the best possible performance

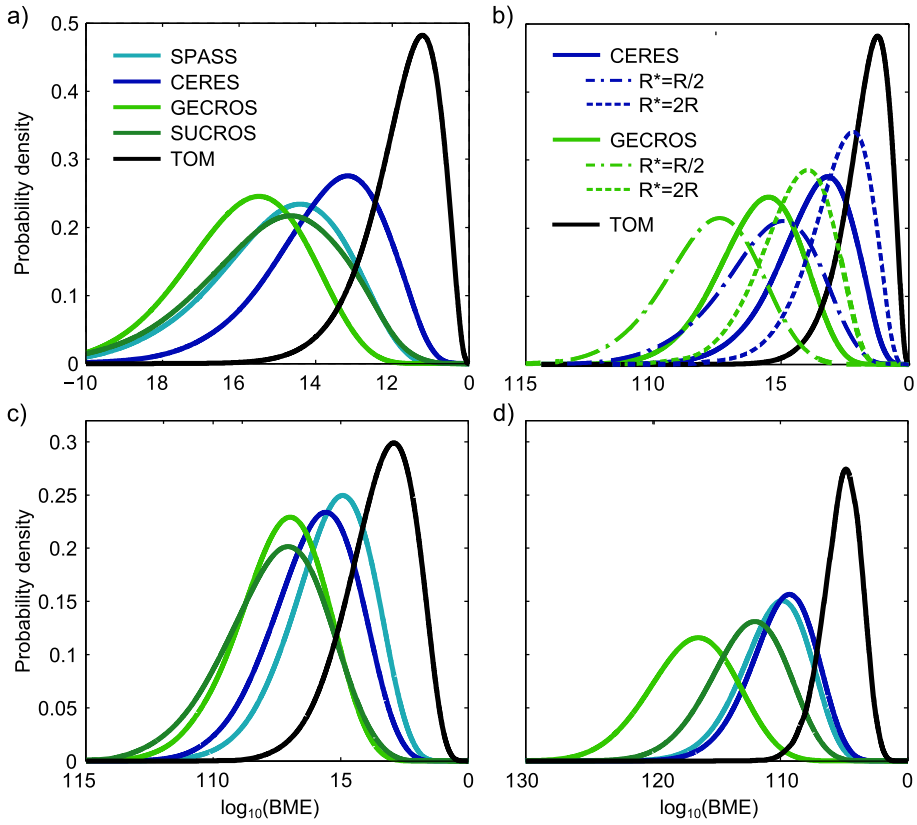


Figure 6. Distributions of $\log_{10}$(BME) obtained by the competing models and by the theoretically optimal model (TOM) based on (a) calibration scenario 1 (using LAI data) with realistic measurement error variance, (b) calibration scenario 1 with reduced and increased measurement error variance, (c) calibration scenario 2 (using ET_a data), and (d) calibration scenario 3 (using both LAI and ET_a data).

as represented by the theoretically optimal model (TOM). This optimal performance is limited by the presence of measurement noise in the calibration data set. In Figure 6a, we show the resulting probability densities of log-BME for all four competing models as well as for the TOM based on calibration scenario 1 (using LAI data).

The relatively large overlap between the distributions of log-BME for CERES with the theoretically optimal distribution (TOM) is reassuring, because it indicates that the model achieves a BME performance score very close to the actual upper limit. This can be interpreted as evidence that CERES provides adequate predictions, given the uncertainty in observing the available system response (measurement noise in the LAI data). In the case of reduced measurement noise, the distance in performance between the TOM and the individual models grows. This is exemplarily shown in Figure 6b for the best and worst performing models CERES and GECROS, respectively. The opposite is the case for increased measurement noise, i.e., it is more difficult to distinguish the models from the TOM because of the low information content in the measurements. Also, the distance between the competing models is reduced.

The resulting probability densities of log-BME for the four competing models and for the TOM based on calibration scenarios 2 (using ET_a data) and 3 (using both LAI and ET_a data) are shown in Figures 6c and 6d, respectively. Note that the distribution of the TOM widens, because the upper limit of performance is a function of data set size (see section 2.5), and data set size increases with each calibration scenario (8 LAI data points, 16 ET_a data points, and 24 data points in the case of both data types). As already discussed with regard to the variability in model weights, the different calibration scenarios reveal different overlaps between the models. When calibrating on ET_a data, e.g., SPASS now shows the largest overlap with the distribution of the TOM, and GECROS and SUCROS show a significant overlap with each other.

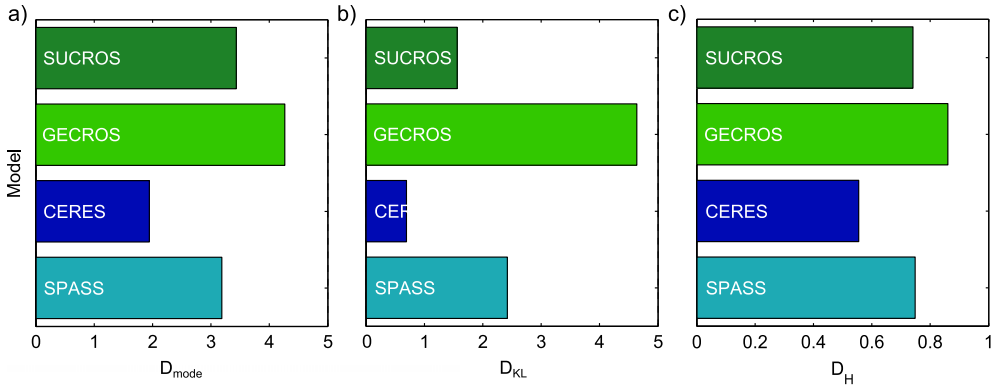


Figure 7. Comparison of $\log_{10}$(BME) distributions (cf. Figure 6) obtained by the four models with the $\log_{10}$(BME) distribution obtained by the theoretically optimal model (TOM) based on calibration scenario 1 (using LAI data). (a) Distance between the modes, (b) Kullback-Leibler divergence, and (c) Hellinger distance.

The remaining distance or room for model improvement between the individual models and the TOM can be measured through the distance between the modes of the distributions (Figure 7a), through the KL divergence (Figure 7b), or through the Hellinger distance (Figure 7c) as discussed in section 2.6. Due to the respective changes in the distributions under varied measurement noise (Figure 6b), those distances increase with decreased measurement error variance, and decrease with increased measurement noise. Note that the TOM does not appear in this figure because its distance to itself would be zero. The distance measures are exemplarily shown here for calibration scenario 1, and could be analogously calculated for the other calibration scenarios.

The different distance measures reveal different aspects of model performance. With all three measures, CERES clearly has the lowest and GECROS the largest distance from the optimal performance. However, SPASS and SUCROS show a very similar distance only when looking at the distribution modes and the relative Hellinger distance between the distributions, but show a substantially different KL divergence. Recall that determining the KL divergence involves taking the log-transform of probability densities, which increases the relative importance of small probabilities. Since we analyze probability density functions of log-BME, which in turn emphasizes small BME values, the KL divergence might lead to inconclusive results.

4.6. Comparison of Differences in Performance Between Models and TOM Based on Bayes Factors

As explained in section 2.7, BME values can also be compared directly by using Bayes factors. We determine log-BF values for each combination of BME values based on a specific outcome of measurement noise, and thus capture actual differences in performance (in BME) between the TOM and the individual models. Negative values indicate that the model has obtained a larger BME score for a specific perturbed data set than the TOM, positive values state that the model performs worse. The cumulative distribution functions of log-BF are shown in Figure 8a. Based on these distribution functions, we compare the competing models in the spirit of hypothesis testing. The significance level at which a model can be rejected with decisive evidence corresponds to the cumulative density value that cuts the red line in Figure 8a. It can be seen that the best performing model CERES can only be rejected against the perfect model at a significance level of about 50% (i.e., for about 50% of the analyzed perturbed data sets). The worst performing model GECROS, in contrast, can be rejected with decisive evidence at a significance level of about 6%.

Figure 8b shows the cumulative distribution functions in the case of decreased or increased measurement noise for the best and worst performing models CERES and GECROS, respectively. In line with the related findings presented in section 4.3, a model can be rejected with decisive evidence at a lower significance level if the noise in the data is smaller. If measurement noise plays a larger role, the significance level rises accordingly.

5. Summary and Conclusions

Bayesian model averaging (BMA) ranks competing conceptual models according to their plausibility in light of a specific calibration data set. Model selection or model averaging can then be performed based on the

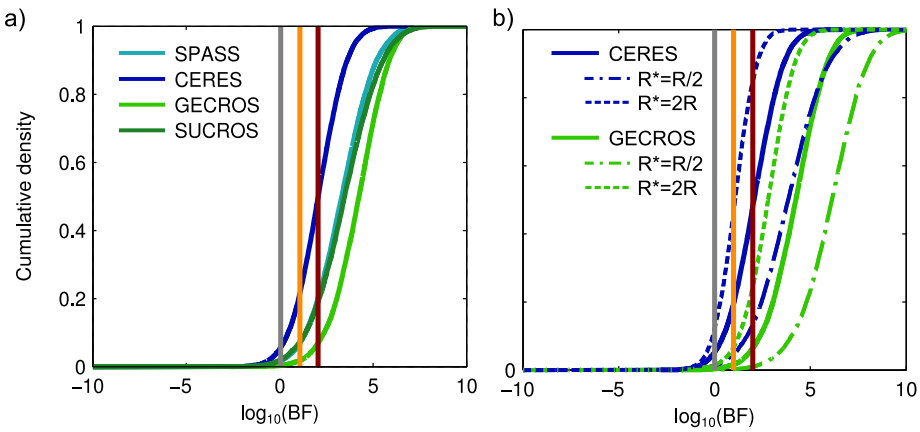


Figure 8. Distributions of $\log_{10}$(Bayes factor) obtained by the competing models in favor of the theoretically optimal model (TOM) based on calibration scenario 1 (using LAI data) with (a) realistic measurement error variance and (b) reduced and increased measurement error variance. Gray lines indicate equally strong evidence for the TOM and for the respective model. Orange and red lines indicate thresholds for strong and decisive evidence (according to *Jeffreys* [1961]) in favor of the TOM against the respective model.

obtained weights. The BMA procedure requires the evaluation of Bayesian model evidence (BME), which is the likelihood of the calibration data averaged over a model's probability space. BME itself is a function of the calibration data and thus, implicitly, also a function of the outcome of measurement error in this specific data set. Acknowledging this dependence, it needs to be assessed whether or not the obtained model weights are robust against measurement noise. Robust model weights are desired, no matter what the goal of applying the BMA approach is—be it model selection or model averaging. In this study, we have focused on model ranking as a direct result of model weights.

If model weights turn out to vary considerably under measurement noise for a specific application, this compromises the reliability in model ranking and triggers the need for data collection techniques with increased accuracy or with a more efficient measurement design. A more efficient measurement design could be found by optimal design strategies [*Atkinson et al*., 2007] that evaluate the benefit of using different data types or data points (in space and time) with regard to an optimality criterion. Each proposed design should be evaluated with regard to the decisiveness in model ranking [e.g., *Atkinson and Fedorov*, 1975] while explicitly considering the impact of random outcomes of measurement noise as proposed in this study.

Not only noise in the calibration data, but also other sources of uncertainty such as uncertain observations or conceptualizations of model input (forcings) or additional stochastic model components turn model weights into uncertain quantities. We have provided a statistical framework that accounts for the resulting variability in model weights from a frequentist perspective. It rests on a brute-force Monte Carlo approach that hosts the BMA routine. Applying the parametric bootstrap method, we resample the uncertain variable (e.g., measurement noise in input or output data or the outcome of a stochastic model for model structural error) by drawing random realizations from its assumed distribution. We then repeat the BMA analysis for every sample and obtain statistics of BME and model weights. These statistics can be analyzed with regard to the confidence in model ranking. For example, determining statistics of the Bayes factor allows to assess the significance (or spread) in pairwise model ranking. It provides a basis for deciding whether a model can be reliably selected or discarded from the set, or whether the considered source of uncertainty for model weights is simply too dominant to make an informed decision.

Our suggested upgrade is general and comes at little additional computational costs, so that the extended BMA routine can be used for an in-depth assessment of uncertainty in model choice in a wide range of disciplines. In the practical application to soil-plant model selection presented here, we have exclusively focused on measurement noise in the calibration data as source of uncertainty for model weights, because an assumption about its statistics is required anyway in the BMA analysis to define the likelihood function. Our results have shown that, in all calibration scenarios investigated here, model weights vary significantly under the impact of measurement noise in the observed data. Since we have not considered model-specific

structural errors in the BMA analysis, it remains an open question for future research how the interplay between structural error models and measurement noise influences the robustness of model weights.

In the specific case of measurement noise in output data, a theoretical limit to model performance exists. We propose to assess a model's distance from this theoretically optimal performance by treating the original unperturbed data set as the theoretically optimal model (TOM). We have discussed various ways to measure this distance and again recommend to use statistics of Bayes factors. Assessing a model's performance in presence of measurement noise with our suggested approach helps to decide, whether more effort should be invested in improving the existing models and/or extending the set with better performing models, or whether, first of all, more accurate data are needed to make differences between the models and the TOM more evident.

Our approach could also be applied to other modeling tasks with significant errors in measurements of driving forces or with uncertainty in boundary conditions. The extended BMA routine could reveal the impact of these additional sources of uncertainty for model weights on model ranking results. The relevance of weighting uncertainty due to input uncertainty remains to be investigated in future case studies.

To sum up, we draw the following conclusions from this study:

1. Uncertainty in model input or output data induces uncertainty in model weights, which we refer to as weighting uncertainty.
2. Weighting uncertainty can severely compromise the confidence in model ranking.
3. Weighting uncertainty needs to be accounted for in a model ranking framework, e.g., by our suggested extension.
4. The extended BMA routine provides a solid basis for data worth analysis and optimal design of experiments toward maximum confidence in model ranking.

Acknowledging the existence of weighting uncertainty triggers the question how this uncertainty is propagated to model-averaged predictions. While in this study, we have proposed a conceptual and numerical framework to reveal the frequentist variability in model weight outcomes, the theoretical extension of the Bayesian total predictive variance formulation to account for weighting uncertainty will be derived in a follow-up study. In particular, we will attempt to resolve the difficulty of mingling frequentist confidence intervals with Bayesian credible intervals.

Appendix A: On the Handling of the Sure Thing Hypothesis in Bayesian Model Averaging

Establishing the observed data set as a model in the spirit of the sure thing hypothesis [*Jaynes*, 2003] is of no use for predictive purposes and therefore intuitively not an adequate competitor to any conceptual model. This seems to contradict the concept of BMA, because one might think that such a model would always win against any competing conceptual predictive model since it obviates the trade-off between bias and variance [e.g., *Burnham and Anderson*, 2003]. The sure thing hypothesis would obtain the maximum possible likelihood value (in this case equal to BME, because there is no probability space to integrate over), which depends only on the shape of the likelihood function (i.e., on the assumed statistical distribution of measurement error). This apparent inconsistency can be resolved as follows: before actually observing the data set (i.e., in *foresight*), there is no way one could come up with the one sure thing hypothesis; instead, there are infinitely many competing hypotheses one would have to consider with the same justification, i.e., the number of measurement points to the power of the number of possible values that the measurements can assume (which is infinity for continuous variables). To have comparable predictive capabilities to competing conceptual models, the model hypothesis thus needs to be rephrased: "There are this many data points to predict, and each of them has that many possible outcomes—the observed data set will be one of those combinations." To account for the multiple alternative subhypotheses, the model prior for the sure thing hypothesis would have to be divided by the total number of equally likely subhypotheses, which will then reduce its posterior model weight below the model weights of any competing conceptual model (in this case actually to zero, since the prior is $1/N_m \cdot 1/\infty = 0$, when assuming equal prior weights $1/N_m$ for all conceptual models). In conclusion, the sure thing hypothesis in *hindsight* is unfair and must not be considered in the BMA competition. The sure thing hypothesis in prediction mode (in *foresight*) has infinite

variance, and would in fact lose any BMA race. In this study, we are only concerned with the BME score that the sure thing hypothesis would achieve given randomly perturbed data sets, and do not include it into the actual model ranking.

Acknowledgments
The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the International Research Training Group "Integrated Hydrosystem Modelling" (IRTG 1829) at the University of Tübingen and within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart. This work was also supported by a grant from the Ministry of Science, Research and Arts of Baden-Württemberg (AZ Zu 33-721.3-2) and the Helmholtz Centre for Environmental Research, Leipzig (UFZ). The authors thank Sebastian Gayler for setting up the soil-plant models for the Nellingen field site. They further acknowledge Hans-Dieter Wizemann and Petra Högy, University of Hohenheim, for providing the experimental data from monitoring campaigns that were funded by the Integrated DFG Project "Regional Climate Change," PAK 346, Germany. Data requests should be directed to the DFG Research Unit 1695 at the University of Hohenheim (contact: https://klimawandel.uni-hohenheim.de/koordination?&L=1). Finally, the authors would like to thank Rodrigo Rojas for insightful discussions on the manuscript and Grey Nearing and two anonymous reviewers as well as the Associate Editor for very constructive comments that helped to strengthen the manuscript and to inspire future research.

References

Abramowitz, G., and H. Gupta (2008), Toward a model space and model independence metric, *Geophys. Res. Lett.*, *35*, L05705, doi:10.1029/2007GL032834.

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

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki, pp. 367–281, Akadémiai Kiadó, Budapest.

Atkinson, A. C., and V. V. Fedorov (1975), Optimal design: Experiments for discriminating between several models, *Biometrika*, *62*(2), 289–303, doi:10.1093/biomet/62.2.289.

Atkinson, A. C., A. N. Donev, and R. D. Tobias (2007), *Optimum Experimental Designs, With SAS, Oxford Stat. Sci. Ser.*, vol. 34., chap. 20, Oxford Univ. Press, Oxford.

Bhattacharyya, A. (1943), On a measure of divergence between two statistical populations defined by their probability distributions, *Bull. Calcutta Math. Soc.*, *35*, 99–109.

Biernath, C., S. Gayler, S. Bittner, C. Klein, P. Hogy, A. Fangmeier, and E. Priesack (2011), Evaluating the ability of four crop models to predict different environmental impacts on spring wheat grown in open-top chambers, *Eur. J. Agron.*, *35*(2), 71–82, doi:10.1016/j.eja.2011.04.001.

Bonan, G. B., K. W. Oleson, M. Vertenstein, S. Levis, X. Zeng, Y. Dai, R. E. Dickinson, and Z.-L. Yang (2002), The land surface climatology of the community land model coupled to the NCAR community climate model*, *J. Clim.*, *15*(22), 3123–3149.

Bowman, A. W., and A. Azzalini (1997), *Applied Smoothing Techniques for Data Analysis: The Kernel Approach With S-Plus Illustrations*, Oxford Univ. Press, Oxford.

Burnham, K. P., and D. R. Anderson (2003), *Model Selection and Multimodel Inference*, chap. 1, 2nd ed., Springer, N. Y.

Carlin, B. P., and T. A. Louis (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, chap. 4, 2nd ed., Chapman and Hall, N. Y.

Carrera, J., and S. P. Neuman (1986), Estimation of aquifer parameters under transient and steady-state conditions: 3. Application to synthetic and field data, *Water Resour. Res.*, *22*(2), 228–242, doi:10.1029/WR022i002p00228.

Davison, A. C., and D. V. Hinkley (1997), *Bootstrap Methods and Their Application*, chap. 4.2, 1st ed., Cambridge Univ. Press.

Draper, D. (1995), Assessment and propagation of model uncertainty, *J. R. Stat. Soc. Ser. B*, *57*(1), 45–97.

Foglia, L., S. W. Mehl, M. C. Hill, and P. Burlando (2013), Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland, *Water Resour. Res.*, *49*, 260–282, doi:10.1029/2011WR011779.

Gayler, S., E. Wang, E. Priesack, T. Schaaf, and F.-X. Maidl (2002), Modeling biomass growth, N-uptake and phenological development of potato crop, *Geoderma*, *105*(3), 367–383.

Gayler, S., J. Ingwersen, E. Priesack, T. Wöhling, V. Wulfmeyer, and T. Streck (2013), Assessing the relevance of subsurface processes for the simulation of evapotranspiration and soil moisture dynamics with CLM3.5: Comparison with field data and crop model simulations, *Environ. Earth Sci.*, *69*(2), 415–427, doi:10.1007/s12665-013-2309-z.

Gayler, S., T. Wöhling, M. Grzeschik, J. Ingwersen, H. D. Wizemann, K. Warrach-Sagi, P. Högy, S. Attinger, T. Streck, and V. Wulfmeyer (2014), Incorporating dynamic root growth enhances the performance of Noah-MP at two contrasting winter wheat field sites, *Water Resour. Res.*, *50*, 1337–1356, doi:10.1002/2013WR014634.

Greve, P., K. Warrach-Sagi, and V. Wulfmeyer (2013), Evaluating soil water content in a WRF-Noah downscaling experiment, *J. Appl. Meteorol. Climatol.*, *52*(10), 2312–2327.

Gull, S. F. (1988), Bayesian inductive inference and maximum entropy, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, vol. 1, pp. 53–74, Kluwer Academic Publishers, Dordrecht, Netherlands.

Hald, A. (1998), *A History of Mathematical Statistics From 1750 to 1930*, chap. 27, John Wiley, N. Y.

Hellinger, E. (1909), Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen, *J. Reine Angew. Math.*, *136*, 210–271, doi:10.1515/crll.1909.136.210.

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

Houska, T., S. Multsch, P. Kraft, H. G. Frede, and L. Breuer (2014), Monte Carlo-based calibration and uncertainty analysis of a coupled plant growth and hydrological model, *Biogeosciences*, *11*(7), 2069–2082, doi:10.5194/bg-11-2069-2014.

Hutson, J. L., and R. Wagenet (1992), *LEACHM: Leaching Estimation and Chemistry Model: A Process-Based Model of Water and Solute Movement, Transformations, Plant Uptake and Chemical Reactions in the Unsaturated Zone; Version 3.0*, *Tech. Rep. Res. Ser.*, vol. 93–103, Cornell Univ., Cent. for Environ. Res., Ithaca, N. Y.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, 296 pp., Cambridge Univ. Press, Cambridge.

Jeffreys, H. (1961), *Theory of Probability*, p. 432, 3rd ed., Oxford Univ. Press, Oxford.

Johnsson, H., L. Bergstrom, P. E. Jansson, and K. Paustian (1987), Simulated nitrogen dynamics and losses in a layered agricultural soil, *Agric. Ecosyst. Environ.*, *18*(4), 333–356, doi:10.1016/0167-8809(87)90099-5.

Kass, R. E., and A. E. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, *90*(430), 773–795, doi:10.2307/2291091.

Kavetski, D., G. Kuczera, and S. W. Franks (2006), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, *331*(1–2), 161–177, doi:10.1016/j.jhydrol.2006.05.010.

Kullback, S., and R. A. Leibler (1951), On information and sufficiency, *Ann. Math. Stat.*, *22*, 79–86, doi:10.1214/aoms/1177729694.

Leube, P. C., A. Geiges, and W. Nowak (2012), Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design, *Water Resour. Res.*, *48*, W02501, doi:10.1029/2010WR010137.

Lu, D., M. Ye, S. P. Neuman, and L. Xue (2012), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, *Adv. Water Resour.*, *35*, 69–82, doi:10.1016/j.advwatres.2011.10.007.

Molz, F. J. (1981), Models of water transport in the soil-plant system: A review, *Water Resour. Res.*, *17*(5), 1245–1260, doi:10.1029/WR017i005p01245.

Morales-Casique, E., S. P. Neuman, and V. V. Vesselinov (2010), Maximum likelihood Bayesian averaging of airflow models in unsaturated fractured tuff using Occam and variance windows, *Stochastic Environ. Res. Risk Assess.*, *24*(6), 863–880, doi:10.1007/s00477-010-0383-2.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, and D. A. Stainforth (2004), Quantification of modelling uncertainties in a large ensemble of climate change simulations, *Nature*, *430*(7001), 768–772, doi:10.1038/Nature02771.

Najafi, M. R., H. Moradkhani, and I. W. Jung (2011), Assessing the uncertainties of hydrologic model selection in climate change impact studies, *Hydrol. Processes*, *25*(18), 2814–2826, doi:10.1002/hyp.8043.

Nearing, G. S., W. T. Crow, K. R. Thorp, M. S. Moran, R. H. Reichle, and H. V. Gupta (2012), Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment, *Water Resour. Res.*, *48*, W05525, doi:10.1029/2011WR011420.

Nearing, G. S., H. V. Gupta, and W. T. Crow (2013), Information loss in approximately Bayesian estimation techniques: A comparison of generative and discriminative approaches to estimating agricultural productivity, *J. Hydrol.*, *507*, 163–173, doi:10.1016/j.jhydrol.2013.10.029.

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

Niu, G., Z. Yang, K. E. Mitchell, F. Chen, M. B. Ek, M. Barlage, A. Kumar, K. Manning, D. Niyogi, and E. Rosero (2011), The community Noah land surface model with multiparameterization options (NoahMP): 1. Model description and evaluation with local-scale measurements, *J. Geophys. Res.*, *116*, D12109, doi:10.1029/2010JD015139.

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

Priesack, E., and S. Gayler (2009), *Agricultural Crop Models: Concepts of Resource Acquisition and Assimilate Partitioning*, pp. 195–222, Springer, Berlin Heidelberg.

Priesack, E., S. Gayler, and H. P. Hartmann (2006), The impact of crop growth sub-model choice on simulated water and nitrogen balances, *Nutrient Cycling Agroecosyst.*, *75*(1–3), 1–13, doi:10.1007/s107500-006-9006-1.

Raftery, A. E. (1995), Bayesian model selection in social research, *Sociol. Methodol.*, *25*, 111–163, doi:10.2307/271063.

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

Refsgaard, J. C., S. Christensen, T. O. Sonnenborg, D. Seifert, A. L. Hojberg, and L. Troldborg (2012), Review of strategies for handling geological uncertainty in groundwater flow and transport modeling, *Adv. Water Resour.*, *36*, 36–50, doi:10.1016/j.advwatres.2011.04.006.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.

Ritchie, J., D. Godwin, and S. Otter-Nacke (1988), *CERES-Wheat: A Simulation Model of Wheat Growth and Development*, Univ. of Tex. Press, Austin.

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

Rojas, R., S. Kahunde, L. Peeters, O. Batelaan, L. Feyen, and A. Dassargues (2010a), Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling, *J. Hydrol.*, *394*(3–4), 416–435, doi:10.1016/j.jhydrol.2010.09.016.

Rojas, R., L. Feyen, O. Batelaan, and A. Dassargues (2010b), On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling, *Water Resour. Rese.*, *46*, W08520, doi:10.1029/2009WR008822.

Schöniger, A., T. Wöhling, L. Samaniego, and W. Nowak (2014), Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence, *Water Resour. Res.*, *50*, 9484–9513, doi:10.1002/2014WR016062.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464, doi:10.1214/aos/1176344136.

Simunek, J., K. Huang, and M. Van Genuchten (1998), The HYDRUS code for simulating the one-dimensional movement of water, heat, and multiple solutes in variably saturated media. Version 6.0, *Tech. Rep. 144*, U.S. Salinity Lab., U.S. Dep. of Agric., Agric. Res. Serv., Riverside, Calif.

Singh, A., S. Mishra, and G. Ruskauff (2010), Model averaging techniques for quantifying conceptual model uncertainty, *Ground Water*, *48*(5), 701–715, doi:10.1111/j.1745-6584.2009.00642.x.

Skilling, J. (2006), Nested sampling for general Bayesian computation, *Bayesian Anal.*, *1*(4), 833–859.

Smith, A. F. M., and A. E. Gelfand (1992), Bayesian statistics without tears—A sampling resampling perspective, *Am. Stat.*, *46*(2), 84–88, doi:10.2307/2684170.

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

Van Genuchten, M. T. (1980), A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Sci. Soc. Am. J.*, *44*(5), 892–898.

Van Laar, H. H., J. Goudriaan, and H. Van Keulen (1997), SUCROS97: Simulation of crop growth for potential and water-limited production situations, Quantitative Approaches in Systems Analysis, No. 14. C.T. de Wit Graduate School for Production Ecology and Resource Conservation, 52 pp./appendices. Wageningen, Netherlands.

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

Wang, E., and T. Engel (2000), SPASS: A generic process-oriented crop model with versatile windows interfaces, *Environ. Modell. Software*, *15*(2), 179–188.

Wizemann, H. D., J. Ingwersen, P. Högy, K. Warrach-Sagi, T. Streck, and V. Wulfmeyer (2015), Three year observations of water vapor and energy fluxes over agricultural crops in two regional climates of Southwest Germany, *Meteorol. Z.*, *24*(1), 39–59, doi:10.1127/metz/2014/0618.

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

Wöhling, T., S. Gayler, E. Priesack, J. Ingwersen, H.-D. Wizemann, P. Högy, M. Cuntz, S. Attinger, V. Wulfmeyer, and T. Streck (2013), Multiresponse, multiobjective calibration as a diagnostic tool to compare accuracy and structural limitations of five coupled soil-plant models and CLM3.5, *Water Resour. Res.*, *49*, 8200–8221, doi:10.1002/2013WR014536.

Wöhling, T., A. Schöniger, S. Gayler, and W. Nowak (2015), Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction, *Water Resour. Res.*, *51*, 2825–2846, doi:10.1002/2014WR016292.

Xue, L., D. Zhang, A. Guadagnini, and S. P. Neuman (2014), Multimodel Bayesian analysis of groundwater data worth, *Water Resour. Res.*, *50*, 8481–8496, doi:10.1002/2014WR015503.

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

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

Ye, M., K. F. Pohlmann, J. B. Chapman, G. M. Pohll, and D. M. Reeves (2010), A model-averaging method for assessing groundwater conceptual model uncertainty, *Ground Water*, *48*(5), 716–728, doi:10.1111/j.1745-6584.2009.00633.x.

Yin, X., and H. van Laar (2005), *Crop Systems Dynamics: An Ecophysiological Model of Genotype-by-Environment Interactions (GECROS)*, Wageningen Acad. Publ., Wageningen, Netherlands.