Water Resources Research

RESEARCH ARTICLE
10.1029/2022WR033711

Information-Theoretic Scores for Bayesian Model Selection and Similarity Analysis: Concept and Application to a Groundwater Problem

Maria Fernanda Morales Oreamuno[1], Sergey Oladyshkin[1], and Wolfgang Nowak[1]

[1]Department of Stochastic Simulation and Safety Research for Hydrosystems, Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, Stuttgart, Germany

Key Points:

- We complement Bayesian model evidence (BME) with information-theoretic scores for Bayesian model selection (BMS) and similarity analysis
- We demonstrate that, unlike BME, relative entropy is suited to compare models that use different subsets of the data for calibration
- We provide a BMS methodology based on Bayesian and information-theoretic scores, including approaches that help to interpret the results

Correspondence to:

M. F. Morales Oreamuno,
maria.morales@iws.uni-stuttgart.de

Citation:

Morales Oreamuno, M. F., Oladyshkin, S., & Nowak, W. (2023). Information-theoretic scores for Bayesian model selection and similarity analysis: Concept and application to a groundwater problem. *Water Resources Research*, *59*, e2022WR033711. https://doi.org/10.1029/2022WR033711

Received 21 SEP 2022
Accepted 8 JUL 2023

Abstract Bayesian model selection (BMS) and Bayesian model justifiability analysis (BMJ) provide a statistically rigorous framework for comparing competing models through the use of Bayesian model evidence (BME). However, a BME-based analysis has two main limitations: (a) it does not account for a model's posterior predictive performance after using the data for calibration and (b) it leads to biased results when comparing models that use different subsets of the observations for calibration. To address these limitations, we propose augmenting BMS and BMJ analyses with additional information-theoretic measures: expected log-predictive density (ELPD), relative entropy (RE) and information entropy (IE). Exploring the connection between Bayesian inference and information theory, we explicitly link BME and ELPD together with RE and IE to highlight the information flow in BMS and BMJ analyses. We show how to compute and interpret these scores alongside BME, and apply the framework to a controlled 2D groundwater setup featuring five models, one of which uses a subset of the data for calibration. Our results show how the information-theoretic scores complement BME by providing a more complete picture concerning the Bayesian updating process. Additionally, we demonstrate how both RE and IE can be used to objectively compare models that feature different data sets for calibration. Overall, the introduced Bayesian information-theoretic framework can lead to a better-informed decision by incorporating a model's post-calibration predictive performance, by allowing to work with different subsets of the data and by considering the usefulness of the data in the Bayesian updating process.

1. Introduction

Environmental modeling allows researchers to reproduce physical systems under different conditions, be they current or future, for design, management or decision making purposes. Due to the high complexity involved in environmental modeling, simplifications and assumptions are necessary to consider the different processes that interact with each other (Wainwright & Mulligan, 2013). Consequently, different sources of uncertainty arise in environmental modeling, including parameter, model input, measurement uncertainty and conceptual uncertainty (Gong et al., 2013; Refsgaard et al., 2007). The latter, also referred to as structural uncertainty, pertains to the choice of model itself, and has gained renewed interest in the past decades as an important source of predictive uncertainty (Bredehoeft, 2005; Gong et al., 2013; Gupta et al., 2012; Höge et al., 2019; Neuman, 2003; Rojas et al., 2008).

Due to incomplete knowledge on the real system, there is not only a single way of representing a given physical phenomenon. Therefore, multiple models can be used to reproduce it, with different levels of detail and complexity (J. Smith & Smith, 2007). Consequently, subjectively limiting the number of possible models to only one can result in an underestimation of the chosen model's uncertainty or in an overconfidence in its predictive capabilities. This, in turn, can lead to biased results, especially with regards to parameter calibration, which could be compensating for errors regarding the model selection (Neuman, 2003; Rojas et al., 2008; Ye et al., 2004).

Therefore, the problem becomes centered around the question of which model to use to represent the true, unknown system, given the current, limited knowledge on it. A widely accepted method to tackle conceptual uncertainty is through multi-model approaches (Bredehoeft, 2005; Neuman & Wierenga, 2003; Refsgaard et al., 2006). Here, a group of competing conceptual models are either generated or selected, and then tested against some acceptance criteria regarding, for example, model fit, model complexity, consistency or multi-objective criteria (Neuman, 2003). Enemark et al. (2019) present a list of publications where conceptual uncertainty in groundwater

systems was considered through multi-model approaches, showing the importance this topic has been given in previous years. Deterministic approaches to multi-model selection use model performance criteria, such as mean square error, Nash-Sutcliffe efficiency (Nash & Sutcliffe, 1970) and cross validation methods (Jung, 2018; Stone, 1974) as model comparison criteria (Gupta et al., 2009). Nevertheless, these do not allow to account for parameter uncertainty. Criteria that do allow to account for parameter uncertainty are stochastic approaches, for example, based on Bayesian theory (Hoeting et al., 1999) or information theory (Gong et al., 2013). These two are the focus of the current paper.

Bayesian multi-model approaches, such as Bayesian model selection (BMS; Raftery, 1995) are based off of Bayes' theorem (Kolmogorov & Bharucha-Reid, 2018). They provide a rigorous stochastic framework to rank and select among competing models, while also considering parameter, input and measurement uncertainty (Draper, 1995). In BMS, a prior belief with regard to model adequacy is updated to posterior model weights, based on observed data. Traditionally, model ranking in the BMS framework is based on the values of Bayesian model evidence (BME), which are defined as the probability of a model of reproducing the available data (Draper, 1995; Raftery, 1995). Such BME-based model selection approaches have been used in many fields for model ranking, and/or selection purposes, for example,: Schöniger, Illman, et al. (2015) and Elshall and Ye (2019) for groundwater modeling, Wöhling et al. (2015) for crop modeling, Marshall et al. (2005) for hydrological models, Brunetti et al. (2017) in hydrogeophysical modeling and Schäfer Rodrigues Silva et al. (2020) in reactive groundwater transport models, to name a few. Additionally, Mohammadi et al. (2018) and Scheurer et al. (2021) apply BMS using surrogate models for sediment transport in rivers and for biochemical processes in the subsurface, respectively. Other Bayesian approaches focus on model averaging, instead of model selection. One case would be Bayesian model averaging (Hoeting et al., 1999), in which the posterior distributions of several models are mixed via a weighted average, with the model weights derived from BME (Höge et al., 2019). However, the current paper focuses on model selection and will therefore not discuss model-averaging approaches.

BME is also referred to as the marginal likelihood, since it is computed by estimating the average of the model likelihood over the entire prior parameter space (Kass & Raftery, 1995). Thus, it often requires multidimensional integration, which can come at high computational costs. Consequently, several approximations for the calculation of BME exist to avoid the previously mentioned integration, including the harmonic mean estimate (Newton & Raftery, 1994), marginal likelihood calculations by Gelfand and Dey (1994) and Chib and Jeliazkov (2001) (see Liu and Liu (2012) for an overview), the Bayesian information criterion (Schwarz, 1978) and the Kayshap information criterion (Kashyap, 1982), to name a few. These, however, require assumptions regarding the posterior distribution, and/or consider point-wise calculations, which can lead to biased results (Schöniger et al., 2014). The Monte Carlo (MC) sampling technique (Hammersley, 1960) provides a bias-free framework to approximate BME, given that it allows to sample from the entire prior parameter space, without additional assumptions about the posterior. In spite of presenting high computational costs, it has shown to provide the best results based on a benchmark test by Schöniger et al. (2014).

In addition to BMS, Schöniger, Illman, et al. (2015) apply a model comparison methodology based not on the true observation data but on an inter-model comparison, and called it Bayesian justifiability analysis (BMJ). In BMJ, each competing model takes turns being the true data generator and is compared against all other models, including itself, in a Bayesian setup. The results, composed of BME-weights, are then summarized in a model confusion matrix (MCM). The term confusion matrix is borrowed from machine learning, where it is used for classification-type problems (see Tharwat, 2020). Similar as with the machine learning application, the MCM allows to visualize similarities between the considered models and to justify model complexity, given the available data. It can therefore complement the model selection analysis. Recently, the BMJ framework has been extended to computationally demanding models applying surrogates (Schäfer Rodrigues Silva et al., 2020; Scheurer et al., 2021) and for model uncertainty quantification (Reuschen et al., 2021).

Even though traditional BME-based BMS analyses do provide a statistically rigorous methodology for considering uncertainties, it does present some limitations, which also extend to the BMJ methodology. A first limitation is that BME does not allow to judge the performance of a model based on its posterior predictive capabilities. Basically, in Bayesia theory (see Gelman et al., 1995; Kolmogorov & Bharucha-Reid, 2018), BME is considered as a normalization factor that can be obtained via the integration of the likelihood over the prior parameter space. It is the average likelihood of a model *before* performing Bayesian update. Therefore, BME values contain only partial information required for the Bayesian updating of a model via the observation data. Additionally, BME

values are highly sensitive to prior selection (Raftery, 1995). In other words, BME-based approaches fail to give an idea of the posterior predictive capabilities or how much the model was able to learn from the data, which are integral steps within the Bayesian framework.

From a model comparison perspective, it is not possible with BME to directly compare models that use different data sets for calibration. This limitation applies when comparing models with different configurations, such as different time/space discretizations, different state spaces, model dimensions (1D, 2D), or models that consider different processes (e.g., flow-only vs. flow-and-transport). For brevity, we will call this multi-fidelity model comparisons. In said cases, simplifications could require averaging available observations or ignoring subsets and/or types of data (e.g., see Mouris et al., 2023). Multi-fidelity comparisons can be useful, not only to select the best model, but also to quantify the change in model performance under different model configurations. Therefore, using BME on these different data sets to compare the models' performance would induce bias. The bias results from the dependence of likelihood functions on data set properties, such as data set size and measurement error. Several studies have addressed the impact of measurement error, data type and data set size on model selection rankings (see Rojas et al., 2010; Schöniger, Wöhling, & Nowak, 2015; Wöhling et al., 2015). They show how one can obtain significantly different BMS weights depending on the data set chosen for the comparison. Thus, a solely BME-based approach is limited to comparing models with the exact same calibration/testing data set to avoid bias in the comparison.

One way to deal with the problems posed by BME is through the use of information theory, which has close ties to Bayesian inference, given that Bayesian inference is linked to maximum entropy quantification and is efficient in terms of information content (Zellner, 1988). Information theory scores include the expected-log posterior density (ELPD), relative entropy (RE), also known as Kullback-Leibler divergence (Kullback & Leibler, 1951), and information entropy (IE), which stem from Shannon's definition of entropy (Shannon, 1948). They have been widely used in probability theory applications to quantify the uncertainty and amount of information (Murari et al., 2019), for model selection purposes (Cliff et al., 2018; Gelman et al., 2014; Murari et al., 2019; Vecer, 2019) and optimal experimental design (Lindley, 1956; Nowak & Guthke, 2016).

Many applications use approximations of entropy, such as the Akaike information criteria (AIC) (Akaike, 1974), Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010), AICc and the multivariate Gaussian posterior estimate (Oladyshkin & Nowak, 2019), due to the difficulty to calculate entropy values for high-dimensional problems. These, as previously mentioned, make use of, assumptions with regards to the posterior distribution, which can cause bias in the results (Oladyshkin & Nowak, 2019). To overcome this, Oladyshkin and Nowak (2019) elaborate on the connection between Bayesian inference and information theory and present prior-based techniques to compute BME in combination with cross entropy (CE), RE and IE. The authors expose the potential benefits of using the additional information criteria to measure information content in Bayesian updating, optimal experimental design and model selection purposes. The paper, however, does not present a specific application. The approach proposed in Oladyshkin and Nowak (2019) has been applied in active learning techniques for surrogate model generation, which closely resembles optimal experimental design setups (Mouris et al., 2023; Oladyshkin et al., 2020), but not, to the authors' knowledge, for model selection or similarity analysis.

The prior-based techniques for estimating RE, CE, and IE proposed in Oladyshkin and Nowak (2019) present some advantages, which can be exploited for model selection and comparison purposes. First, they avoid any additional assumptions and skip any multidimensional integration or density estimation. Second, the information-theoretic scores provide information on the updating process within the Bayesian inference framework, which is ignored in traditional BME-based BMS and BMJ analysis. Furthermore, some of the information-theoretic scores exhibit a reduced dependence on data set properties and remain meaningful for comparing models that use different calibration data sets. Thus, in a Bayesian context, they can be utilized to compare models with different configurations and/or process that use different calibration data sets.

The current paper proposes to complement the traditional BME-based methodology with information-theoretic scores to overcome the two aforementioned limitations surrounding BME. We focus on ELPD as a measure of information between the likelihood and the posterior (posterior model fit), RE between the prior and the posterior (updatability of model parameters through the data) and IE of the posterior for model selection and comparison purposes. To avoid additional assumptions we will use prior-based MC sampling (Gelman et al., 1995; Hammersley, 1960) alongside the formulations presented in Oladyshkin and Nowak (2019). Additionally, and building on the work by Schöniger, Illman, et al. (2015), we implement a model similarity analysis using model

confusion matrices (MCM) based on BME and the different information-theoretic scores, to determine differences and similarities between the models in different steps of the Bayesian updating process.

We apply and test the methodology on a synthetic groundwater model setup, made up of four competing models and based on the setup in Schöniger, Illman, et al. (2015). Here, four flow-transport models with different spatial hydraulic conductivity distributions are compared against a set of synthetically generated data (BMS) and against each other (model similarity analysis). This setup will allow to test our proposed methodology on environmental models with different complexity, represented by their spatial distribution of hydraulic conductivity. Taking advantage of the current setup, we include a fifth model, which will represent a lower-fidelity version of one of the previously mentioned models: a flow-only model, which considers fewer processes. Consequently, it can be calibrated on a smaller data set only. This second case will allow us to test the methodology on models with different calibration data sets.

The setup resembles potential real-world scenarios, where a modeler might be faced with the decision of determining a model's (prior) complexity, which processes to consider or whether to gather additional information. Considering multiple models, as in the proposed setup, can reduce the conceptual uncertainty, allowing to reach a more informed decision. We chose a synthetic, computationally cheap model to exemplify the methodology. Therefore, with this study and its application case, we seek to (a) present the behavior of the information-theoretic scores within the BMS and model similarity frameworks, and how they can be interpreted to complement BME; (b) determine which scores can be used to compare models that use different data sets for calibration, and the limitations associated to them.

The remainder of the paper is organized as follows: in Section 2 we present an overview of traditional BMS and BMJ frameworks. We then introduce the synthetic setup in Section 3. We briefly present the different information scores, as well as a computationally simple way to calculate them in Section 4. Here, we also show how these scores overcome the current limitations of BME-based BMS and BMJ approaches and we guide the reader in how to interpret them within both frameworks. Lastly, the results and discussion are presented in Section 5.

2. Bayesian Model Assessment Framework

2.1. Bayes' Theorem

In Bayes' theorem (see Kolmogorov & Bharucha-Reid, 2018), current knowledge associated with the set of uncertain parameters, for a given model M_k, is encoded in a so-called prior distribution. The current beliefs are then updated based on how well the model can reproduce observed data to obtain a posterior distribution (Raftery, 1995). Bayes' theorem can be summarized by the following equation:

$$p(\boldsymbol{\omega}_k|M_k,\mathbf{y}_o)=\frac{p(\mathbf{y}_o|\boldsymbol{\omega}_k,M_k)\,p(\boldsymbol{\omega}_k|M_k)}{p(\mathbf{y}_o|M_k)}, \tag{1}$$

where $p(\boldsymbol{\omega}_k|M_k)$ is the prior distribution of modeling parameters $\boldsymbol{\omega}_k$ from the parameter space Ω_k, $p(\mathbf{y}_o|\boldsymbol{\omega}_k, M_k)$ is the likelihood function, $p(\boldsymbol{\omega}_k|M_k, \mathbf{y}_o)$ is the updated posterior distribution and the denominator $p(\mathbf{y}_o|M_k)$ is the probability of data given M_k. The latter could be seen as a normalizing factor to obtain the posterior distribution and is referred to as BME.

The likelihood function (see Aldrich, 1997) incorporates the available observation and quantifies model M_k's fit to the available observation data y_o (Press, 2009). If one assumes Gaussian-distributed independent errors, as we do for the purpose of this paper, a multivariate Gaussian distribution can be used as a likelihood function:

$$p(y_o|\boldsymbol{\omega},M_k)=(2\pi)^{\frac{-N_o}{2}}|\mathbf{R}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{y}_k-\mathbf{y}_o)^T\mathbf{R}^{-1}(\mathbf{y}_k-\mathbf{y}_o)\right], \tag{2}$$

where $\mathbf{R}$ is the (here diagonal) covariance matrix of measurement errors of size $N_o \times N_o$, with N_o being the number of observations in the calibration data set, $\mathbf{y}_o$ is the vector of calibration data (observations) and $\mathbf{y}_k$ is the vector of corresponding model results from model M_k. The term to the left of the exponent is a normalizing factor, such that the area under the likelihood function integrates to one over the distribution of measurement error. The goodness of fit to the data is encoded in the exponential term on the right. Extended approaches exists that account for auto-correlated and/or non-Gaussian errors, or that include statistical representations of model

inaccuracies. As we will not exploit specific properties of Equation 2, our assumption does not induce any loss of generality.

The equation for BME can be rewritten as follows:

$$p(\mathbf{y}_o|M_k) = BME_k = \int_{\Omega_k} p(\mathbf{y}_o|\boldsymbol{\omega}_k, M_k)\, p(\boldsymbol{\omega}_k|M_k)\, d\boldsymbol{\omega}_k, \tag{3}$$

or, shortly, using the prior-based expectation $\mathbb{E}_{prior}[\cdot]$:

$$BME_k = \mathbb{E}_{prior}\left[\, p(\mathbf{y}_o|\, \boldsymbol{\omega}_k, M_k)\,\right], \tag{4}$$

where the BME value is expressed as an integral over the total parameter space Ω_k and, for that reason, also known as the marginal likelihood (Kass & Raftery, 1995). Based on this formulation, BME values are sensitive to prior selection (Kass & Raftery, 1995) and, therefore, tend to favor models with the best compromise between model flexibility and model fit (Schöniger, Illman, et al., 2015). There are several alternative approaches to estimate BME using posterior marginalization or additional approximations (see Oladyshkin & Nowak, 2019; Schöniger et al., 2014). However, Equation 3 is often employed using the prior-based brute MC sampling (Hammersley, 1960), where BME is estimated as an average of the prior-based likelihoods.

It is well-known that MC sampling to estimate BME requires a large number of model realizations (N_{MC}) and can therefore become computationally prohibitive. Nevertheless, this sampling technique, compared to others, avoids additional assumptions with regards to posterior distributions and point-wise estimations (see details in Schöniger, Illman, et al. (2015) and Oladyshkin and Nowak (2019)). Therefore, in the current paper, we follow the MC sampling strategy to avoid additional assumptions and biased results.

2.2. Bayesian Model Selection

In a similar manner as with parameter uncertainty, Bayes' theorem can be used to quantify conceptual uncertainty associated to model choice through BMS. Here, both the prior parameter and model adequacy beliefs of model M_k are updated based on the observed data to obtain posterior parameter distributions and posterior model weights (Chipman et al., 2001). Considering a finite number of competing models N_M, the BMS formulation for a given model M_k can be summarized by the following equation (Hoeting et al., 1999):

$$W(M_k|\mathbf{y}_o) = \frac{p(\mathbf{y}_o|M_k)\, W(M_k)}{\sum_{i=1}^{N_M} p(\mathbf{y}_o|M_i)\, W(M_i)}, \tag{5}$$

where $W(M_k)$ and $W(M_k|\mathbf{y}_o)$ are the model prior and posterior weights associated to a given competing model M_k, respectively. The use of a uniform distribution of $1/N_M$ is often used as a prior model assumption, since it allows for the updated model weight to depend solely on the model's fit to the data, and not on subjective prior distributions (Chipman et al., 2001; Press, 2009). The denominator in Equation 5 is a normalizing factor and is the same across all models. Therefore, the only term that has an effect on the posterior model weight is $p(\mathbf{y}_o|M_k)$, which is the BME for model M_k and quantifies the goodness of fit of model M_k against the available data.

As BME is a relative measure of model fit associated to a model, a strategy for model selection is to choose the model with the highest posterior model weight (Chipman et al., 2001; Oladyshkin & Nowak, 2019), given that a higher BME indicates the best compromise between the model fit and the model's flexibility. Model flexibility can be described in terms of variance in the output values due to the parameter distributions. Usually, wide, uninformative distributions associated to sensitive parameters result in a wide range of output values. BME values are valid only for the current state of knowledge, and are dependent on the data. Additionally, the resulting model weights depend on the set of models being analyzed. This implies that, if more knowledge is gained on the real values (additional measurements) or additional models are considered, the BMS weights (W) will generally change.

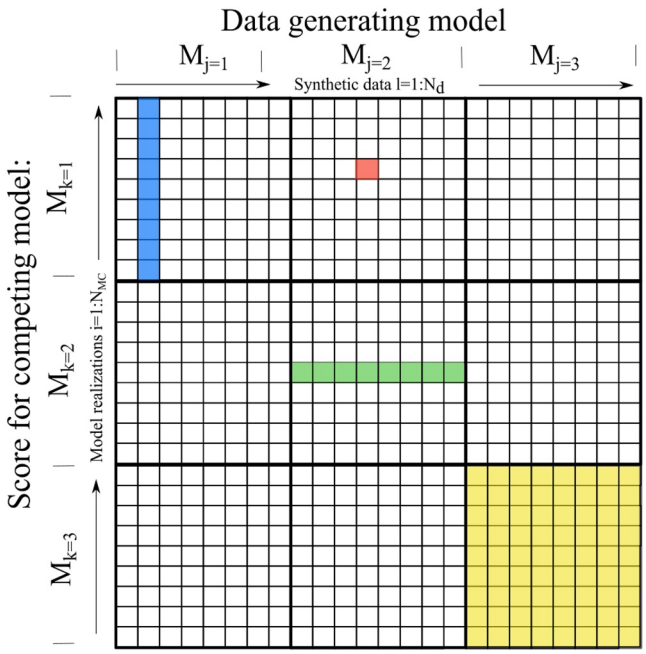

Figure 1. Schematic illustration how to construct a model confusion matrix. Red box: likelihood of a single realization drawn from model $M_{k=1}$, given a single realization drawn from model $M_{j=2}$. Blue boxes: average likelihood (Bayesian model evidence (BME)) of model $M_{k=1}$ given a single realization of model $M_{j=1}$. Green boxes: Average BME values for model $M_{k=2}$ given all realization of model $M_{j=2}$. The diagonal boxes (e.g., yellow box) correspond to the average $\overline{BME}_k^{(j)}$ for a model M_k given data-generating model M_j.

2.3. Bayesian Justifiability Analysis

In a BMJ setup, as applied by Schöniger, Illman, et al. (2015), the goal is to test whether the complexity (e.g., parameter number and spread of their prior) of models would be justifiable when facing a limited data set, under the assumption that the models could actually be true. To this end, the models are not compared against observed data (as in BMS) but against each other, in a synthetic setup. Here, each competing model takes turns in being the data-generating model and is then compared against all competing models, including itself, within the Bayesian modeling framework. As a result, BME weights are obtained for each data-generating/competing model combination.

In BMJ, N_d realizations from the parameter prior of each data-generating model M_j are sampled and evaluated in the model. Noise is then added to each data set to account for the measurement error associated to real observations (Reuschen et al., 2021). Each model data set $\mathbf{y}^*_{l,j}$, with $l = 1 \ldots N_d$, then takes turns being the "true" data for model M_j, and the Bayesian framework is applied for each competing model M_k. The BME weights $\left(BME_k^{(j)}\right)$ are then averaged over all N_d realizations, to obtain an averaged $\overline{BME}_k^{(j)}$ value, as summarized by the following equation:

$$\overline{BME}_k^{(j)} = \frac{1}{N_d} \sum_{l=1}^{N_d} BME_{k,l}^{(j)}, \tag{6}$$

where $\overline{BME}_k^{(j)}$ is the averaged BME of model M_k given N_d realizations of model M_j. The results for all $\overline{BME}_k^{(j)}$ are then summarized in a so-called MCM (Schöniger, Illman, et al., 2015). The MCM has the size $N_M \times N_M$, where the columns represent the data-generating models, M_j, and the rows represent the competing models M_k. Confusion matrices, also referred to as contingency or error matrices, are often used in machine learning applications, for example, classification problems (see Lindholm et al., 2022).

Figure 1 shows a schematic illustration of generating the MCM. Following the order set by Equation 6, each $\mathbf{y}^*_{l,j}$ (each column in Figure 1) takes turns in being the true data and the Bayesian framework is applied for each competing model M_k. The red-highlighted box in Figure 1 represents the likelihood value obtained when comparing each individual realization $M_{k,i}$ of model M_k for $i = 1 \ldots N_{MC}$, against a single synthetic data set $M_{j,l}$ generated by model M_j. Expectation over N_{MC} realizations of the competing model M_k is schematically displayed by each row in Figure 1 (blue highlighted boxes), which results in $BME_{k,l}^{(j)}$. The averaged weights for each realization of model M_k given $M_{j,l}$ are represented by the entries along the green cells in Figure 1. Lastly, these $BME_{k,l}^{(j)}$ values are averaged to obtain the MCM entries, represented by the yellow area in Figure 1.

Similar to the confusion matrices in classification problems, the diagonal values in the resulting MCM correspond to how much a model measures up against itself as the data generator, while the off-diagonal values correspond to how the models measure up against each other. Therefore, diagonal weights close to 1 indicate that the model can identify itself as the true data generator, and does not confuse its results (Schöniger, Illman, et al., 2015). On the other hand, diagonal values close to $1/N_M$ indicate that a model confuses its predictions with those of other models. This can be caused by either models being very similar in their predictions, or the available data set size not being big enough for a model to identify itself (Schöniger, Illman, et al., 2015). Therefore, "the [MCM] reveals whether two models are actually very similar in their predictions, while the conventional BMS analysis cannot distinguish this case from the case of two models that by chance achieve a similar overall goodness of fit" (Schöniger, Illman, et al., 2015). A similar type of analysis, but with the main focus on off-diagonal values was used by Schäfer Rodrigues Silva et al. (2020) to reveal and discuss similarities within a set of models.

2.4. Effect of Different Calibration Data Sets

BME's dependence on properties of calibration data sets, such as data set size, informativeness and measurement error, comes from the likelihood function in Equation 2. When the same data set is used for all competing models, the normalization factor $(2\pi)^{\frac{-N_o}{2}}|\mathbf{R}|^{-1/2}$ in Equation 2 is the same for all models and therefore cancels out when estimating Bayesian model weights. In this case, the effect of data set properties is concentrated inside the exponential term, in the values and size of $\mathbf{R}$, where it divides the difference between modeled and observed data, or model fit.

Canceling the normalization factor is not possible if models with different data sets (including different data set sizes and/or measurement errors) are considered, for example, in a multi-fidelity model comparison, as discussed in Section 1. Thus, the effect of data set properties on the normalization factor must be taken into account and will directly affect the BME value, independent on how well the model reproduces the observations. Indeed, from Equation 2, one can see that the first term, $(2\pi)^{\frac{-N_o}{2}}$, decreases with increasing data set size, decreasing likelihood and BME values. In the second term, $|\mathbf{R}|^{-1/2}$, the value of the determinant of $\mathbf{R}$ depends on both data set size and on the magnitude of the measurement error. Consequently, BME becomes biased when comparing models that make use of different subsets of the available data set for calibration/testing, since the models would not be tested under equal conditions. We will further explore this scenario using a groundwater case study, in which we compare models that use different subsets of a calibration data set. We will use this example to expose the problems with BME in these cases, and how we can use information-theoretic scores to overcome them. We describe the groundwater case study in Section 3, followed by a description of the information-theoretic scores in Section 4.

3. Description of Groundwater Case Study

There is a high uncertainty associated to subsurface modeling, especially with regard to spatially variable material parameters and the different processes involved (James & Oldenburg, 1997). Therefore, there is not a unique conceptual/mathematical representation of such systems that satisfies all applications. This topic has been tackled in many studies, including in Schöniger, Illman, et al. (2015) and Rojas et al. (2008), to name a few. As in Schöniger, Illman, et al. (2015), we focus on the conceptual uncertainty associated to the spatial characterization of hydraulic conductivity. Different parametrizations, with different complexities, represent different models, which can be compared in a BMS setup to select the best one, given a set of available data. Loosely based on the work by Schöniger, Illman, et al. (2015), we build a synthetic groundwater model setup, where the challenge of modeling subsurface heterogeneity is examined by comparing four models with different spatial distributions of hydraulic conductivity in a Bayesian context. The four competing models are flow-and-transport models, and are compared against the same calibration data set.

As previously mentioned, modelers can be interested in a multi-fidelity comparison of models, not only with the goal of selecting the best one, but also to characterize their behavior under different configurations, at different scales or dimensions or when considering different processes. This can provide insight into the influence and utility of additional observations, processes and/or complexities on a model's output. Therefore, taking advantage of the available setup, we build an additional model that only considers flow processes, using one of the spatial hydraulic conductivity parametrization. Said model can only work with the flow-related subset of the calibration data set, and therefore cannot be compared under traditional BMS setups to its flow-and-transport counterpart. To overcome this challenge, we propose to test this additional scenario under different calibration data sets using the combined Bayesian and information-theoretic methodology.

Consequently, the following five models are considered in our setup:

1. Transport homogeneous model (hm)
2. Transport zoned model with five zones (zm_5)
3. Flow zoned model with five zones (zm_{5_f})
4. Transport zoned model with nine zones (zm_9)
5. Transport geostatistically distributed model (gm),

Table 1
Boundary Conditions and Constant Aquifer and Transport Parameters

Parameter	Value
Domain size	[50, 50 m]
Grid size	[1, 1 m]
West BC[a]	1 m
East BC[a]	0 m
North BC[a]	0 m/s
South BC[a]	0 m/s
Porosity	0.35
Longitudinal dispersivity	2.5 m
Transverse dispersivity	0.5 m
Diffusion coefficient	$1{\cdot}10^{-9}$ m/s

[a]BC, Boundary condition.

Where first the four transport models are compared against the same data set in a BMS setup. Second, the two 5-zoned models with different calibration data sets are compared using additional Bayesian information-theoretic scores.

We generate the observation data sets using a synthetic run of one of the competing models, as opposed to an experimental laboratory setting, as was the case in Schöniger, Illman, et al. (2015). This provides a controlled setup, where we know beforehand both the synthetically true observations and the synthetic, true hydraulic conductivity distribution.

Through this application, we seek to demonstrate the behavior of the additional Bayesian information-theoretic scores for (a) models with different conceptual representations (prior flexibility) and the same calibration data set and (b) multi-fidelity representations of models, using different calibration data sets. We also plan to address how the Bayesian information-theoretic scores can be used to assess similarities between models through MCMs. To do so, we will first summarize the simulation setup as well as the competing models in the current section, followed by the results obtained for both the model selection and model similarity analysis in Section 5.

3.1. Synthetic Groundwater Model Setup

For generating the groundwater models, we use a MATLAB-based finite element method code, based on the program used in Schöniger (2010). The program solves the steady state, 2D groundwater transport equations for a 50 m × 50 m confined aquifer, discretized every 1 m. A Dirichlet boundary condition of 1 and 0 m were set on the west and east boundaries, respectively, and impermeable Neumann boundary conditions were assigned to the north and south boundaries. Additionally, a tracer plume was located in the middle of the west boundary, with a fixed boundary concentration of $c_{in} = 1$ over a length of 10 m. For all competing models, the boundary conditions and the different transport parameters were kept constant. The model constants are summarized in Table 1. More information on the model setup can be found in Schöniger (2010) and Nowak and Cirpka (2006).

We consider four different hydraulic conductivity (K) models to generate five groundwater models, following the logic presented in Schöniger, Illman, et al. (2015). We consider the parameters in the log scale, to avoid negative conductivity values. Therefore, we will refer to the hydraulic conductivity as $\ln(K)$. The homogeneous model represents the simplest model, since it consists of a single $\ln(K)$ value assigned to all cells in the grid. We consider two zoned models, one divided into five homogeneous $\ln(K)$ zones, and one divided into nine, with the latter therefore being more flexible. For these three models, we assume that the independent $\ln(K)$ per zone values follow a normal distribution with a mean of $\ln(1{\cdot}10^{-5})$ and a variance of 1. Lastly, the most flexible model is represented by an isotropic geostatistical model, in which $\ln(K)$ follows a multivariate Gaussian distribution with an exponential covariance function, with a mean of $\ln(1{\cdot}10^{-5})$, a variance of 1 and correlation length of [10, 10 m]. This results in 2,500 uncertain parameters, which are all dependent on each other. A summary of the different $\ln(K)$ parametrization models can be seen in Table 2.

Table 2
Summary of Hydraulic Conductivity Parametrization Models

Model	Number of parameters	Parameters' distribution
Homogeneous (hm)	1	$\mathcal{N}\left[\ln\left(1 \cdot 10^{-5}\right), 1\right]$
5-zoned (zm_5)	5	$\mathcal{N}\left[\ln\left(1 \cdot 10^{-5}\right), 1\right]$
9-zoned (zm_9)	9	$\mathcal{N}\left[\ln\left(1 \cdot 10^{-5}\right), 1\right]$
Geostatistical (gm)	2,500	$\mathcal{N}\left[\ln\left(1 \cdot 10^{-5}\right), \mathbf{\Sigma}\right]$[a]

[a]Σ = Exponential covariance function, with correlation length (x, y) = [10, 10 m].

For this test case, the synthetically true $\ln(K)$ distribution was generated from a realization of the geostatistical model, shown in Figure 2. The synthetic setup and the synthetic observation data generated from it will be discussed further in Section 3.2. To define the extent and location of each discrete, homogeneous zone, we did an informed zonification based on the true $\ln(K)$ distribution to simulate a prior knowledge of the real $\ln(K)$ field. We then sampled independent $\ln(K)$ values for each one, and therefore did not consider a spatial correlation between zones. Both zone classifications can be seen in Figure 3.

We evaluate the model outputs in five, arbitrarily located observation wells within the study area, which are shown in Figure 2. We take the four previously mentioned $\ln(K)$ models as flow-and-transport models, with hydraulic

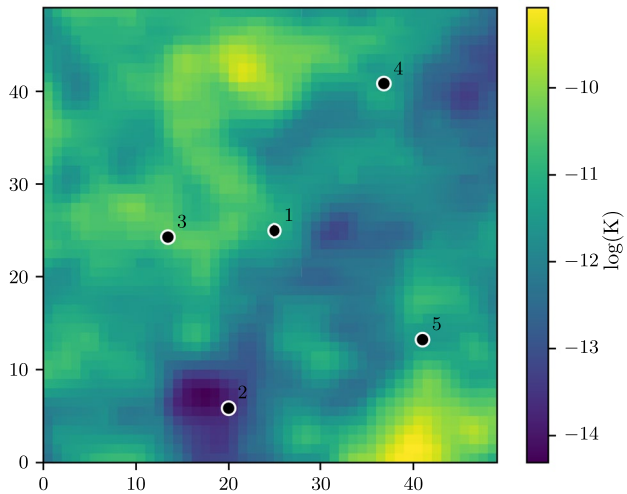

Figure 2. True ln(K) spatial distribution, synthetically generated through the geostatistical model. The black dots correspond to the location of the measurement points.

head (h) and concentration (c_o) measurements. Thus, they count with a calibration data set size of 10. To include a model with a different data set size, we additionally consider the 5-zoned model as a flow-only model, which only considers hydraulic head observations and thus has a calibration data set size of five.

It is worth noting that the homogeneous model represents a unique test case. Despite variations in hydraulic conductivity values, the model generates deterministic hydraulic head values under the given pressure boundary conditions. Furthermore, the model's transport behavior exhibits minimal change, leading to low variability in concentration values across the domain. Consequently, parameter input has little impact on the output values, rendering the homogeneous model insensitive to updates based on the data analyzed in this paper. Nonetheless, the model provides valuable insights into the behavior of the Bayesian and information-theoretic scores, which will be discussed in future sections.

3.2. Synthetic Setup

For the controlled setup, we use a random realization of the geostatistical model as the synthetic, true observed data, since it represents the most flexible model, from both a number of parameters and an output space perspective. The true spatial h and c_o distribution can be seen in Figures 4a and 4b. These represent the data that the competing models will be compared against in a BMS setup.

If one had an infinite number of model realizations, the geostatistical model would be able to reproduce data generated from itself perfectly. To properly account for measurement noise in this synthetic setup for BMS and BMJ analysis, noise was added to the synthetic data set, to account for measurement error (Reuschen et al., 2021). For the noise, we consider a standard deviation of $h_{error} = 0.06$ m and $c_{error} = 0.06 + 20\%$ of the measured concentration (c_o), assuming a relative error for c_o dependent on the measured value.

4. Bayesian Information-Theoretic Model Assessment Framework

The topic of information theory, in the context of communication theory, was addressed by Shannon (1948), and has paved the way to information theory in the context of probability and statistics. More information on the development of information theory can be seen in the works by Kullback (1997) and Commenges (2015), to name a few. This field focuses on quantifying of the amount of information needed to account for uncertainty, referred to as IE. Originally, information theory was introduced for discrete-valued random variables (Shannon, 1948) and then expanded to continuous distributions. Differences with regards to discrete and continuous entropy are further

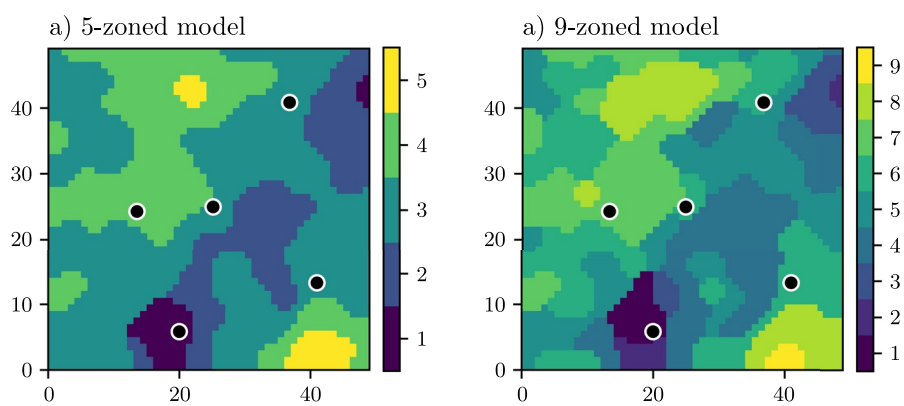

Figure 3. Zone classification for (a) 5-zoned model and (b) 9-zoned model, based on synthetically true ln(K) distribution.

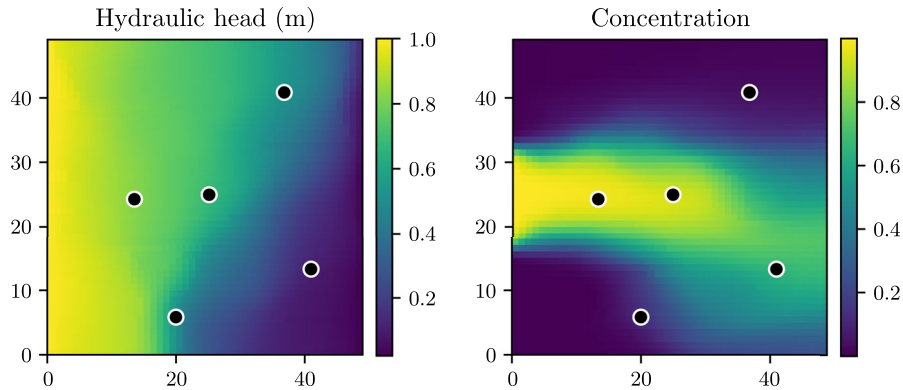

Figure 4. Spatial distribution of hydraulic head (left) and concentration (right) for the true synthetic run, generated with the geostatistical model.

detailed in Marsh (2013) and Santamaría-Bonfil et al. (2016). In the current work, we will explore the connection between information theory for continuous distributions and Bayesian inference as presented in Oladyshkin and Nowak (2019) to enhance the BMS and BMJ concepts presented in Section 2.

We begin with a brief overview of information-theoretic scores, including IE, CE, and RE within the Bayesian framework. This is followed by a computationally simple way to calculate and interpret them within both BMS and BMJ frameworks.

4.1. Definitions of Information-Theoretic Scores

Information entropy describes the quantification of the expected uncertainty, or the missing information required to remove uncertainty from a random variable (Shannon, 1948). In the context of Bayesian theory, the IE of a parameter set $\boldsymbol{\omega}_k$ can be calculated for its prior or posterior probability distribution. In this work, we limit ourselves to quantifying the IE for the posterior to determine the remaining uncertainty after updating the prior based on the observed data. IE of the posterior is formulated as follows:

$$IE \equiv H\left[p(\boldsymbol{\omega}_k \mid M_k, \mathbf{y}_o)\right] = -\int_{\Omega} p(\boldsymbol{\omega}_k \mid M_k \mathbf{y}_o) \ln\left[p(\boldsymbol{\omega}_k \mid M_k, \mathbf{y}_o)\right] d\boldsymbol{\omega}_k, \tag{7}$$

where $H[\cdot]$ is the entropy according to

$$H[p(x)] = -\int p(x) \ln[p(x)]\, dx. \tag{8}$$

CE (Shannon & Weaver, 1949) quantifies the expected missing information to get one distribution from another (Good, 1956; Shore & Johnson, 1980). For the Bayesian framework, one can calculate the information needed to get the posterior $p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o)$ from the prior $p(\boldsymbol{\omega}_k, M_k)$ as follows:

$$CE \equiv H\left[p(\boldsymbol{\omega}_k \mid M_k, \mathbf{y}_o),\ p(\boldsymbol{\omega}_k \mid M_k)\right] = -\int_{\Omega} p(\boldsymbol{\omega}_k \mid M_k, \mathbf{y}_o) \ln[p(\boldsymbol{\omega}_k \mid M_k)]\, d\boldsymbol{\omega}_k, \tag{9}$$

where $H[\cdot, \cdot]$ is the general CE according to

$$H[p(x), q(x)] = -\int p(x) \ln[q(x)] dx. \tag{10}$$

Similar to CE in Equation 9, the expected missing information to get the posterior from the likelihood could also be assessed using a non-normalized cross entropy (NNCE) (Oladyshkin & Nowak, 2019):

$$NNCE \equiv \hat{H}\left[p(\boldsymbol{\omega}_k \mid M_k, \mathbf{y}_o),\ p(\mathbf{y}_o \mid \boldsymbol{\omega}_k, M_k)\right] = -\int_{\Omega} p(\boldsymbol{\omega}_k \mid M_k, \mathbf{y}_o),\ \ln\left[p(\mathbf{y}_o|\boldsymbol{\omega}_k, M_k)\right] d\boldsymbol{\omega}_k. \tag{11}$$

The NNCE is non-normalized since the likelihood is considered a proper probability distribution with respect to the measurement errors for which the likelihood is determined, and not with respect to the model parameters (Oladyshkin & Nowak, 2019). If one eliminates the negative sign in Equation 11, the formulation can be reinterpreted as the expected log-predictive density (ELPD) (see Gelman et al., 2014; Vehtari & Ojanen, 2012), given that the integral in Equation 11 represents a posterior-based expectation of the log-likelihood:

$$ELPD = -NNCE. \tag{12}$$

ELPD, in its different approximations, has been used to compare and quantify model fit based on posterior predictive capacities within the Bayesian framework (Gelman et al., 2014; Höge et al., 2019; Schöniger et al., 2014). It can be used to describe the accuracy with which a model can predict not only the data used for calibration, but also all potential other data, including those used for testing or those not even available yet (Gelman et al., 2014; Nicenboim et al., 2021).

Another score used to compare two probability distributions in terms of uncertainty is RE, also known as the Kullback-Leibler divergence (D_{KL}). Kullback and Leibler (1951) mention that this term can be used as a measure of how different two distributions are, or the amount of information needed to discriminate between them. Various authors remark that RE may seem like a measure of distance between two distributions, since RE ≥ 0 and RE $= 0$ only if both distributions are the same. Nevertheless, it is not a proper measure of distance (Commenges, 2015) since it is not symmetric and thus $RE[A, B] \neq RE[B, A]$. In the Bayesian context, we will use RE to assess the expected gain, or reduction in uncertainty, in going from the prior to the posterior as follows:

$$RE \equiv D_{KL}\left[\, p(\boldsymbol{\omega}_k \mid M_k, \mathbf{y}_o),\ p(\boldsymbol{\omega}_k \mid M_k) \,\right] = \int_{\Omega} \ln\left[\frac{p(\boldsymbol{\omega}_k \mid M_k, \mathbf{y}_o)}{p(\boldsymbol{\omega}_k \mid M_k)}\right] p(\boldsymbol{\omega}_k \mid M_k, \mathbf{y}_o)\, d\boldsymbol{\omega}_k. \tag{13}$$

Using Equations 7 and 9, Equation 13 can be also rewritten as the difference between the CE and the IE for the posterior (IE). In other words, it can be calculated by removing the uncertainty of the posterior from the amount of information needed to get the posterior from the prior:

$$RE = CE - IE. \tag{14}$$

4.2. Computation of Information-Theoretic Scores

Various problems arise when solving Equations 7, 11, and 13. This includes the estimation of the multidimensional integral through additional assumptions, that become necessary in high dimensions (Oladyshkin & Nowak, 2019). In the current paper, we use the following approaches, in order to avoid any assumptions and still excluding multidimensional integration.

4.2.1. ELPD

To compute ELPD (and therefore NNCE), we use a brute-force MC methodology. Given that the posterior parameter and output distributions are usually not known in analytical form, Equation 11 can be rewritten as a sample-wise expectation of the posterior (giving equal weights to each posterior sample):

$$ELPD = \mathbb{E}_{post}\left[\ln\left[p(\mathbf{y}_o|\boldsymbol{\omega}_k, M_k)\right]\right], \tag{15}$$

where $\mathbb{E}_{post}[\cdot]$ is the posterior-based expectation. Additionally, posterior samples are a by-product of Bayesian updating. Therefore, one can approximate Equation 15 by:

$$ELPD \approx \frac{1}{N_{post}} \sum_{i=1}^{N_{post}} \ln[\, p(y_o|\boldsymbol{\omega}_i, M_k)\,], \tag{16}$$

where N_{post} is the total number of posterior parameter sets. Posterior samples can be obtained, for example, through Markov Chain Monte Carlo (MCMC) techniques or via a rejection sampling technique (A. Smith & Gelfand, 1992).

One can observe a similarity between Equation 15 for ELPD and Equation 4 for BME: they are both measurements of model fit, with the former being marginalized on the posterior and the latter on the prior parameter

distribution. Therefore, as with BME, the best model from this perspective is the one with the largest ELPD. In contrast to BME, ELPD has a smaller influence from the prior, given that it does not play a significant role in posterior predictions when having informative data (Gelman et al., 2014). Thus, models with different prior flexibility, which received different BME scores, can receive a similar ELPD value if their posteriors present a similar model fit.

4.2.2. Relative Entropy

In order to compute RE, Oladyshkin and Nowak (2019) reformulate Bayes' theorem from Equation 1 and obtain the following formulation (see Oladyshkin & Nowak, 2019):

$$RE = -\ln[BME] + ELPD. \tag{17}$$

Equation 17 indicates that RE can be calculated based on BME and ELPD (−NNCE), using simple MC for the former and either MCMC or rejection sampling techniques for the latter, as mentioned in the previous sections. Moreover, one can clearly see that the information gained through the data, in the form of RE, is the difference between the prior model fit (through −ln(BME)), and the posterior model fit (through ELPD). From a Bayesian perspective, the model with the largest RE is the one that reduces predictive uncertainty the most when moving from the prior to the posterior parameter distributions, or to which the available data was the most useful. Another way of interpreting RE, as mentioned by Oladyshkin and Nowak (2019), is that a maximum RE is assigned to the model whose overall normalized likelihood function is most similar to the true unknown posterior distribution. This makes RE different yet still suitable as a model selection criterion. The difference is that RE is often inversely related to BME and so can lead to different model selection outcomes. The inverse relation arises as data can be easily informative for a-prior-uninformed models.

4.2.3. Cross Entropy

The CE between the prior and posterior distributions in Equation 9 can be obtained from its definition using the posterior-based expectations (similar to ELPD):

$$CE = -\mathbb{E}_{post}\left[\ln p(\boldsymbol{\omega}_k \mid M_k)\right] \tag{18}$$

or, numerically, using posterior-based sampling:

$$CE \approx -\frac{1}{N_{post}} \sum_{i=1}^{N_{post}} \ln\left[p(\boldsymbol{\omega}_{k,i} \mid M_k)\right]. \tag{19}$$

4.2.4. Information Entropy

With knowledge on ELPD, CE, and RE, one can calculate IE in the Bayesian context directly from Equation 14:

$$IE = CE - RE, \tag{20}$$

employing Equations 19 and 17.

As previously mentioned, IE is the uncertainty associated to the posterior distribution. Consequently, from a model selection perspective, one would be inclined to select the model with the smallest IE (smallest uncertainty). A small IE can be due to (a) a large gain in information by moving from the prior to the posterior and/or (b) a small uncertainty associated to the prior parameter distribution (simple or very informative prior). Another way to interpret IE is through the two components in Equation 20. From the equation we see that IE depends on the difference between CE and RE. Both terms represent different aspects of the relationship between the prior and posterior distributions: RE represents the gain in information when moving from a prior to a posterior distribution and CE represents the uncertainty carried from the prior to the posterior (CE). Therefore, it is important to consider how much of the posterior uncertainty is due solely to the prior (CE), and how much is due to the informativeness of the data (RE) to make an informed decision based on IE. We will further expand on this in Section 5.2.

4.3. Effect of Different Calibration Data Sets

Recall from Equation 16 that ELPD is a likelihood-based score and as such can also lead to biased results when comparing models that use different subsets of available data for calibration, including (a) subsets with different number of observations and/or (b) same subset size but different data types, with different measurement errors. As with BME, the normalizing factor in the likelihood function from Equation 2 cannot be canceled out in said cases. This can be seen in more detail in Equation A4, where the equation for ELPD is decomposed to mathematically see the effect of the normalizing factor from the likelihood function. Therefore, in spite ELPD providing useful information in traditional BMS setups, it should not be used when comparing models that use different calibration data sets.

In contrast to BME and ELPD, RE, and IE scores compare models based on the prior and/or posterior parameter distributions and not directly on model fit: RE quantifies the gain in information from prior to posterior and IE the uncertainty associated to the posterior parameter distribution. Therefore, RE and IE do not depend directly on the likelihood function, and thus are not affected by models with different data set size. This can be seen in Equation 21:

$$\begin{aligned} RE \;=\; & -\ln(\mathbf{NF}) - \ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) + \\ & \ln(\mathbf{NF}) + \mathbb{E}_{post}\left[\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right] \end{aligned} \tag{21}$$

where NF stands for the normalizing constant. Here we show how, when estimating RE, the normalizing factor NF from the likelihood function, present in both ln(BME) and ELPD, is canceled out. Consequently RE, and by definition IE, depend solely on the exponential term of the likelihood function in the prior and posterior parameter spaces, which is a direct measure of model predictive quality. Through this equation, we show that we do not use a different equation for models with different calibration data sets, but take advantage of the existing benefits of a Bayesian approach to estimating RE. We show a more detailed derivation in Equation A5. Due to this, RE and IE are more suitable scores to compare models with different data sets, compared to BME or ELPD.

4.4. Extension of Bayesian Model Selection and Model Similarity Analysis

Based on the additional Bayesian information-theoretic scores presented above, we now update the BMS and BMJ analysis to include said scores. This allows to compare and rank models not only from a prior BME perspective, but also from the perspectives of posterior and information gain. In the case of BMS, calculating ELPD, RE and IE does not require additional computationally expensive calculations, given that they are a direct result of calculating BME (using a MC approach, which is the most computationally demanding step) and the rejection sampling process, intrinsic to the Bayesian framework.

In the case of BMJ analysis, the goal of this paper is not necessarily to justify a model's flexibility (as in the original paper by Schöniger, Illman, et al. (2015)), but to simply compare the models from different perspectives. Therefore, we will refer to it as a model similarity analysis once the information scores are included in the analysis. We propose to construct the MCM for each score in a similar way as for the BME-weights in BMJ (Section 2.3). Hence, additional to BME, we evaluate all information-theoretic scores for each model M_k, given each realization M_l from the data-generating model M_j. To estimate the entries in the MCM, we average each score for all realizations $M_k|M_{j,l}$ (entries along the green cells in Figure 1) as detailed by the following equations:

$$\overline{ELPD}_k^{(j)} = \frac{1}{N_d}\sum_{l=1}^{N_d} ELPD_{k,l}^{(j)}, \tag{22}$$

$$\overline{RE}_k^{(j)} = \frac{1}{N_d}\sum_{l=1}^{N_d}\left(-\ln\left[BME_{k,l}^{(j)}\right] + ELPD_{k,l}^{(j)}\right), \tag{23}$$

$$\overline{IE}_k^{(j)} = \frac{1}{N_d}\sum_{l=1}^{N_d}\left(-RE_{k,l}^{(j)} - CE_k^{(j)}\right). \tag{24}$$

Four MCMs, one for each BMS score, will be generated from the results. We additionally propose to represent BME in the natural logarithmic scale (ln(BME)), so the results are also in terms of entropy and comparable to all

other scores. Nevertheless, its interpretation is the same with or without the log-scale. Therefore, the ln(BME) confusion matrix entries are calculated as follows:

$$\overline{\ln[BME]}_k^{(j)} = \frac{1}{N_d}\sum_{l=1}^{N_d} \ln\left[BME_{k,l}^{(j)} \right], \tag{25}$$

In contrast to Schöniger, Illman, et al. (2015), we do not calculate Bayesian model weights, since these can only be obtained from BME. In contrast, we propose to generate a normalized MCM, where each score for M_k given $M_{j,l}$ is divided, or normalized, by the diagonal value ($k = j$ for each realization l). The goal is to make the results easier to interpret, as opposed to building the MCMs with the direct results from Equations 22–25. The diagonals in the final, normalized MCM will always be equal to 1 and the off-diagonals will indicate how much model M_k diverges, on average, from the diagonals. The closer the normalized value is to 1, the more similar the model M_k is to the data-generating model M_j, given the current state of knowledge. The normalization must be done for each realization $M_{j,l}$ individually and then averaged over all values for M_j (green row in Figure 1).

It is important to highlight that the limitations associated to BME and ELPD extend to the model similarity analysis. Therefore, we should not build BME and ELPD confusion matrices to compare models that use different subsets of available data for calibration. In our current case, the exclusion is due to the flow model being able to manage a smaller subset of the data. However, as previously mentioned, this is not exclusive to models that use calibration sets of different size, but extends to data sets with different data types, that have different associated errors, and thus influence the scores differently. This would be the case if, for example, we sought to compare a flow-only model with a concentration-only model, with the goal of quantifying the effect of each data set individually on a higher-fidelity flow-*and*-transport model.

4.5. Interpretation of MCMs

Based on the explanation of the different scores, the MCMs for ln(BME) and ELPD can be considered as likelihood-based comparisons. As in traditional BMJ, the off-diagonal entries can be interpreted as how well model M_k can reproduce the results from M_j, or how much M_j confuses its results in the prior (ln(BME)) and posterior states (ELPD). We compare the models on a column-wise basis, where the closer the off-diagonal values are to the diagonal, the more similar they are according to the corresponding criterion.

The MCM for RE represents how much each model M_k can learn from observations generated by model M_j. Given that the MCM is built by first evaluating the scores for one realization of $M_{j,l}$ at a time and then averaged, we do not expect the RE values in the diagonal to tend to zero (when the data-generating model is compared against itself). Therefore, two models can be considered similar from a RE perspective if they undergo similar information gains (similar updatability), which would result in off-diagonal normalized values close to 1.

IE confusion matrices represent a posterior-based comparison, quantifying the remaining uncertainty in the posterior of M_k after updating prior beliefs with data from M_j. As per the definition of IE in Equation 20, model similarity based on this perspective depends on a balance between similarities in the prior distribution and updatability based on the data generated by model M_j. Therefore, its interpretation is directly linked to both terms.

5. Illustrative Application to Groundwater Flow and Transport Models

5.1. Numerical Implementation

In this section, we apply the extended BMS and model similarity analysis to the groundwater problem presented in Section 3. We calculate the BME, ELPD, RE, and IE scores using Equations 4, 16, 17, and 20. For this, we sample $1{\cdot}10^6$ MC realizations from each of the five models. For BMS (Section 5.2) we use a single synthetic data set from the geostatistical model to generate the synthetic observations, as described in Section 3.2. For implementing the model similarity analysis (Section 5.3) Equations 22–25 are used to populate the MCMs for ln(BME), ELPD, RE and IE. Here, N_d = 1,000 MC realizations of each possible data-generating model M_j are sampled and then compared to the $N_{MC} = 1{\cdot}10^6$ MC realizations from each competing model M_k. The noise added to the data-generating models is based on the measurement error variances presented in Section 3.

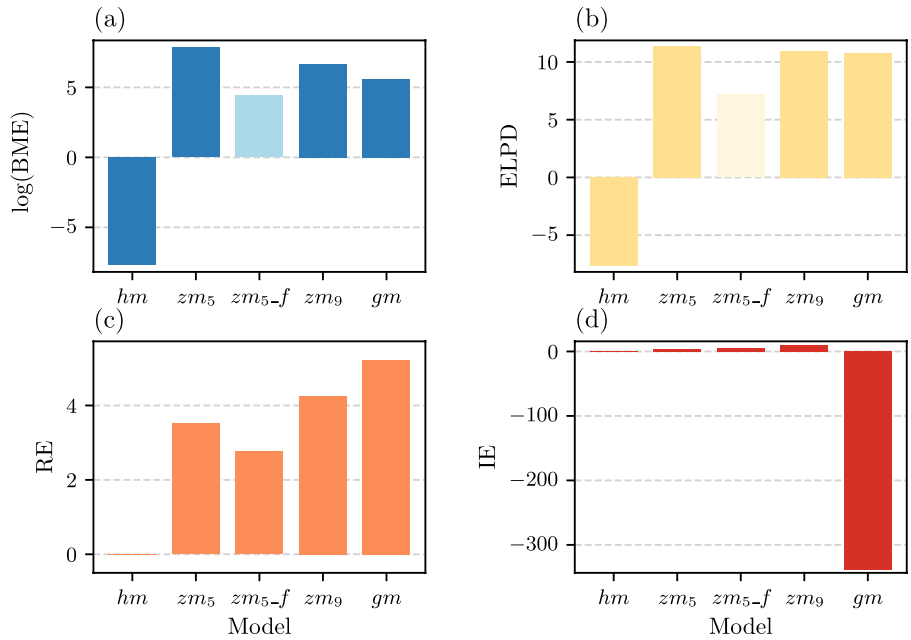

Figure 5. Bayesian and information-theoretic scores for each competing groundwater model in the Bayesian model selection setup. (a) ln(BME), (b) expected log-predictive density, (c) relative entropy, (d) information entropy. The bars with the lighter hues in (a) and (b) indicate that the corresponding flow model cannot be compared to all other models in the set using ln(BME) and ELPD, respectively, given that the model uses a subset of the data for calibration.

We will show, through this application, how to interpret the information-theoretic scores alongside BME. Additionally, recall that we use one model with a different calibration data set (5-zoned flow model) for a multi-fidelity comparison with its flow-and-transport counterpart, which uses a different calibration data set. Here, we show how the likelihood-based scores (BME and ELPD) can no longer be applied and one must switch to a solely RE and IE-based Bayesian comparison.

5.2. Bayesian Model Selection

The Bayesian and information-theoretic scores for BMS analysis can be seen in Figure 5, comparing all models to a random realization of the geostatistical model. We use the results for (a) a comparison between the four transport models from different perspectives, using all four scores and (b) a multi-fidelity comparison between the flow-only and flow-and-transport models, using only RE and IE. Since BME and ELPD cannot be used to compare models that use different subsets of the data for calibration, the scores for the flow model are shown in a lighter hue in Figures 5a and 5b.

5.2.1. BME-Based Selection: Maximization of Data Probability

According to BME, the model with the highest value presents the best compromise between model fit and model flexibility, and would therefore be selected. The results in Figure 5a show that the 5-zoned model obtains a significantly higher value among the four transport models. The homogeneous model, although it has the lowest prior flexibility, receives the overall smallest BME value, indicating an overall bad model fit. Therefore, it would be discarded in a BME-based analysis. The geostatistical model is punished due to a more flexible prior, and thus receives a smaller BME than the 5-zoned model. These results are in line with traditional BME analysis, where the less flexible models are rewarded with a higher score, if they present a good overall model fit. However, as has been pointed out in Section 2, BME does not use the posterior. Consequently, the analysis in Figure 5a can be considered incomplete from a fully Bayesian standpoint, as it considers only fractional information from the entire Bayesian inference.

5.2.2. ELPD-Based Selection: Maximization of Posterior Likelihood

In contrast to BME, the model with the highest ELPD is considered as having the best *posterior* fit and would therefore win against models with a lower score. From Figure 5b, one would favor any among the 5-zoned, 9-zoned and geostatistical model over the homogeneous model given the latter's substantially smaller ELPD (worst overall posterior model fit). The three favored models present similar ELPD scores of 11.46, 11.09, and 10.65, respectively. This means that all three models have posterior parameter distributions that can similarly predict the observed values. This illustrates how ELPD is less dependent on prior choice when compared to BME. The 5-zoned model, however, still presents a slightly higher ELPD score among all four competing models, and would therefore be selected from a posterior perspective.

In this case, the ELPD score serves to support the BME-based decision in favor of the 5-zoned transport model, given that it received the highest BME and the slightly higher ELPD. If, on the other hand, the more flexible geostatistical model had received a significantly higher ELPD, one might want to weigh the additional computational cost associated to Bayesian updating for a more high-dimensional parameter space against a better posterior fit, especially when acknowledging that one will as of now work with posterior models anyways. This proves how ELPD can be used to complement BME by considering a posterior model fit in the decision process, reducing the influence of a potentially uninformative prior choice. However, similar to BME values, the ELPD considers only partial information from a Bayesian inference perspective. Namely, ELPD omits the information gain from prior to posterior (the change in model fit) and, hence, the analysis in Figure 5b can still be considered as incomplete.

Although it is not recommended to compare models that can handle different subsets of the available data for calibration using BME and ELPD, it is possible to observe the direct effect of data set properties on these scores. If we compare the BME and ELPD values for the two 5-zoned models (with and without transport), we can observe that the transport model, which uses all available data for calibration, has higher scores than the flow model. This can be explained by the relatively small measurement error associated to the additional concentration data and the ability of the model to reproduce the true c_o values in both the prior and posterior. However, since the transport model needs to reproduce a larger data set, it is more difficult to achieve high scores. Nevertheless, in this case, the likelihood function rewards the (few) realizations that are able to reproduce all 10 observations within the error threshold with a significantly higher likelihood, increasing the expected BME and ELPD values. This indicates how the measurement error and the size of data set can play an important role when calculating BME and ELPD, which can result in bias when comparing models with different calibration data sets.

5.2.3. RE-Based Selection: Maximization of Relative Information Gain

As previously mentioned, RE presents two main advantages as a model comparison criterion: (a) it allows a combined prior-posterior analysis, considering how useful the data was to the model and (b) it allows to compare models in a multi-fidelity scenario, where the models use different subsets of the data due to different configurations and/or processes being considered. Larger RE values are associated to a high gain in information from the available data. This suggests a greater reduction in a model's predictive uncertainty when moving from a prior to a posterior state.

Focusing first on the results for the multi-fidelity comparison between the two 5-zoned models shown in Figure 5c, we can observe that the transport model obtained a higher RE than the flow-only model. This means that the transport model, by adding an additional process and therefore additional data types, was able to learn more than the flow model from only 5 hydraulic head observations. Therefore, adding the process proved useful to the model's learning process. However, we can observe that the additional gain in information is small, in relation to the flow model's RE. This suggests that most of the gain in the flow-and-transport model is due to the hydraulic head data. From a decision-maker's perspective, these results can give an idea of what kind of data would be most useful to a model, or which processes are better reproduced by the high-fidelity model, from an information gain perspective.

Although it might seem obvious and intuitive, that more data means a better fit, that might not necessarily be the case. A decrease in RE when adding additional processes/observations can indicate a problem with the conceptual model, and therefore one might want to reconsider the approach used for the additional processes. Changes in RE values when adding/removing certain processes can also indicate which models are better suited to reproduce different types of data.

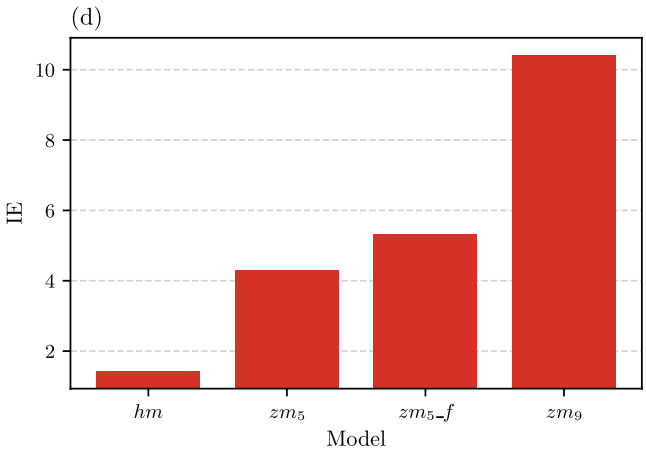

Figure 6. Information entropy scores for a subset of the competing models, excluding the geostatistical model.

Analyzing the BMS results for the transport models, we observe that the geostatistical model obtained the highest RE value. This indicates that it was able to learn the most from the measurement data when updating from the prior to the posterior. This suggests an inverse relationship between RE and BME, given that the geostatistical model also obtained the smallest BME when compared against the zoned models. We found that RE values are not strongly related to model complexity, but rather reflect how well the models are able to learn from the data. In some situations, the ability of a model to learn from the measurement data could coincide with a greater model flexibility, but the latter should not be seen as a necessary or sufficient condition.

Moreover, a small RE does not necessarily indicate a bad fit to the measurement data, but may also be the result from an initially good prior fit. This seems to be the case for the 5-zoned transport model. The smaller RE associated to it can be explained due to an overall good prior fit to the data (small difference between ln(BME) and ELPD). In other words, it had little to learn from the data given that the prior parameter distribution encompassed the true posterior quite well. This can be seen as a limitation when comparing models solely based on RE, given that it tends to punish models with a good prior fit with a smaller score. On the other hand, Figure 5c shows that the homogeneous model obtains a RE value close to 0, which can also be attributed to BME and ELPD presenting similar values. In this case, however, the BME and ELPD scores do present the smallest values and thus the RE score can be interpreted as the homogeneous model not being able to learn from the data due to an overall bad model fit. This can be attributed to the small sensitivity of the homogeneous model's outputs to different parameter values, as mentioned in Section 3.1. Consequently, the model was not able to learn from the available observations. Therefore, we would like to emphasize to the reader that BME, ELPD, and RE can complement each other (when possible) and allow to rank and select among models based on different perspectives or goals of the analysis.

5.2.4. IE-Based Selection: Minimization of Posterior Uncertainty

Recalling from Section 4.2, IE is the uncertainty associated to the posterior state. Consequently, one would be inclined to select a model with a smaller posterior uncertainty. However, IE depends on the interaction between RE and CE. Thus, it is important to consider both the effect of the informativeness of the data through RE and the effect of the prior distribution through CE.

When analyzing the IE results in Figure 5d, we can observe how the geostatistical model presents a significantly smaller IE score than the other four competing models. This would incline us to choose the geostatistical model, given that it would provide the most certain posterior distribution. Nevertheless, if we analyze IE together with RE, we can see that the difference between the RE values (Figure 5c) is not as substantial as that between IE values. We can therefore conclude from Equation 20 that the large difference in IE is due to the prior uncertainty through the CE, and not necessarily due to a greater gain in information from the data. This suggests a substantial influence of the prior distribution on the posterior uncertainty of a model, and how it can overshadow the effect of the data and the overall model fit represented by RE.

Additionally, the smaller IE value associated to the geostatistical model can be attributed to a very informative prior, given the 2,500 uncertain parameters associated to this model and the high correlation between them. Due to these factors, the space where all parameters are within the allowed prior variance is very small. This causes each parameter set to have a high probability density associated to it, which translates to a small entropy (small uncertainty). It is worth mentioning that, if the correlation between the parameters were to substantially decrease, the entropy would increase, given that entropy is maximized for increasingly independent parameters. The opposite happens when the parameters are independent, as in the case of the homogeneous and the zoned models: the probability density associated to each realization decreases with a higher parameter dimension (given parameter independence), and thus the entropy increases.

If we omit the geostatistical model for visualization purposes, as displayed in Figure 6, the homogeneous model presents the smallest IE within the remaining subset. Here, IE indicates that the homogeneous model has a lower posterior parameter uncertainty than the 5 and 9-zoned models. This, however, can be attributed to the prior

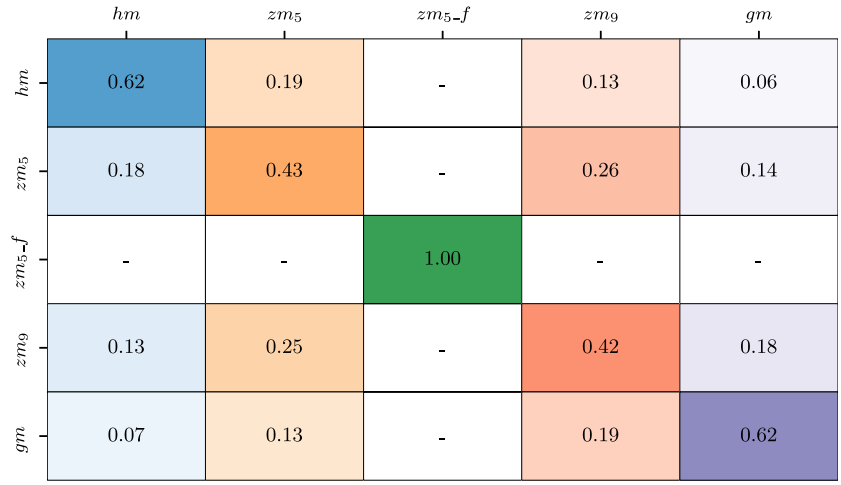

Figure 7. Model confusion matrix based on Bayesian model evidence weights. Columns correspond to data-generating models M_j and rows to competing models M_k.

distribution (more specifically to the number of uncertain parameters) and not to the overall Bayesian updating process. This is supported by the small BME, ELPD, and RE scores associated with the homogeneous model. Additionally, the 9-zoned model received the highest IE score within this subset of models, in spite of having the largest RE among them. It is clear, then, that the IE depends on both the prior parameter uncertainty and how useful the data is in eliminating the uncertainty specified by this prior. However, the dependence of CE on the number of parameters generates biased results when comparing models with different parameter dimensions and should therefore be avoided in such cases.

Similar to RE, IE can also be used to compare the two 5-zoned models, which have the same prior assumptions but use different subsets of the data for calibration. Here, the IE results highlight the effect of additional data on the posterior uncertainty, decreasing the influence of the prior (CE). Figure 6 shows how the flow-and-transport model has the smaller IE among them, suggesting it learned slightly more from the additional processes and managed to reduce the overall posterior uncertainty. This coincides with the results obtained through RE, given that it is the main contributor to the IE score.

5.3. Bayesian Model Similarity Analysis

5.3.1. BME and ELPD: Likelihood-Based Comparison

To analyze the similarities, or differences, between the transport models in their prior states, one could limit oneself to the original BMJ analysis based on BME-weights, which is presented in Figure 7. Here, the rows and columns related to the lower-fidelity flow model are left blank, given that BME cannot be used to compare among models that use different calibration data sets.

From the results in Figure 7, we can observe that both the homogeneous and the geostatistical model receive high diagonal values, indicating their ability to identify their own results. They also have the smallest off-diagonal values, meaning they do not tend to confuse their results. From this, one can conclude that these two models are the most different from each other and from the zoned models. On the other hand, the 5-zoned and the 9-zoned models obtain model weights smaller than 50% on the diagonal, as well as similar off-diagonal values when the respective other is the data generating model. This suggests that these models have the highest likelihood of confusing their results, and thus are the most similar from a prior perspective.

Results from the extended model similarity analysis, as detailed in Section 4.4, are shown in Figure 8. We focus on the off-diagonal values, namely how much they deviate from the behavior of the data-generating model (diagonals). The results are presented as normalized MCMs based on all four scores, including the ln(BME) values. Similar to the BME weights, the ln(BME) Figure 8a and ELPD Figure 8b MCMs show empty rows and columns where the 5-zoned flow model is involved. The non-normalized version of the MCMs can be seen in Figure B1.

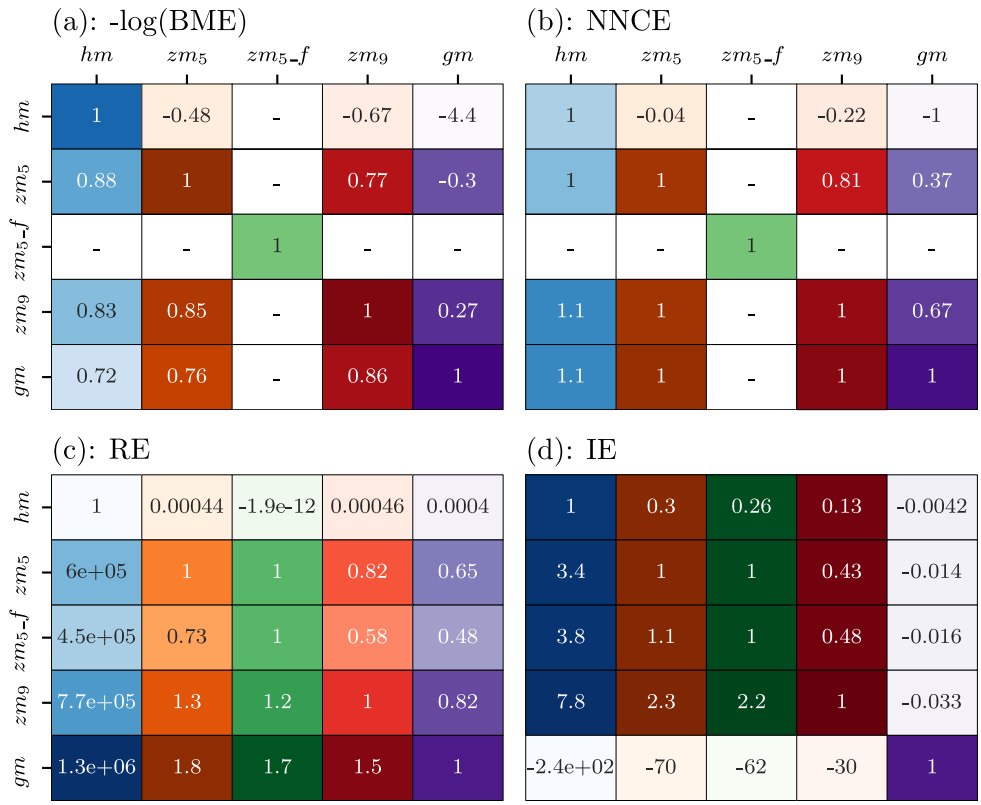

Figure 8. Normalized model confusion matrices for (a) ln(BME), (b) ELPD, (c) relative entropy and (d) information entropy for the Bayesian model similarity analysis. The off-diagonal values are normalized based on the diagonal values and therefore an off-diagonal value equal or close to 1 indicates a large similarity. The more a value diverges from 1, either larger or smaller, the greater the difference between the models, based on the corresponding criteria. The empty cells in (a) and (b) correspond to the cases where the flow model, which uses a smaller calibration data set, is involved.

As with the BME-weight-based MCM, the ln(BME)-based MCM in Figure 8a compares model outputs from a prior predictive perspective. Therefore, one can see similar trends in both results: the homogeneous and the geostatistical model receive the smallest off-diagonal entries when they generate the data, confirming them as the two most different ones. Additionally, the 5 and 9-zoned models obtain the most similar off-diagonal values (closer to 1) when the respective other generates the data, meaning they present similar prior predictive capabilities. One must keep in mind, though, that rescaling values to a log-scale compresses the differences given in a linear scale at large values, and thus the level of similarity based on BME appears different compared to ln(BME). Nevertheless, the trend is maintained and one can reach similar conclusions in terms of model selection and similarity.

In contrast to ln(BME), the ELPD-based MCM in Figure 8b compares models from a posterior predictive perspective. This means, how likely model M_k's posterior predictions are to coincide with those of M_j. The results in Figure 8b show that for all models, except the homogeneous model, the off-diagonal values are closer to 1 than for the prior-based ln(BME) results. This indicates that the models appear more similar in the posterior predictive state than they do in the prior. For example, the differences between the 5 and 9-zoned models seem to have been reduced, given that the off-diagonal values are closer to 1 when the respective other is generating the data. For the geostatistical model, the normalized values along row 5 in Figure 8b are close to one. This, however, does not indicate a larger similarity between the models, given that the same cannot be observed along the last column, when the geostatistical model generates the data. Therefore, it is important to consider both sides of the diagonal to be able to determine similarities between models based on these scores.

5.3.2. RE: Combined Prior and Posterior Comparison

In the case of the RE-based MCM, the rows and columns associated to the lower-fidelity flow model are not blank, given that RE allows for a comparison between models, regardless of the subset of the data being used. However, we focus our comparison, first, between the flow-and-transport models (which use all available measurements for calibration) and second, in a separate analysis, between the two 5-zoned models.

The RE results in Figure 8c allow to compare the models based on a combined prior and posterior perspective, given that it evaluates the updatability of the prior based on the data generated by M_j. Apparently, the homogeneous model is the most different in terms of updatability. The off-diagonal entries along column 1 (when the homogeneous model is M_j) show averaged RE values that are orders of magnitude larger than 1, which suggests that all other models are capable of learning more from data generated by the homogeneous model, than the homogeneous model from itself. This is due to the low sensitivity of the homogeneous model's output to different $\ln(K)$ values. When the homogeneous model acts as M_j, the random noise added to the data-generating model's outputs injects the previously lacking variability to the model, which still cannot be reproduced by the almost deterministic homogeneous model (M_k). Therefore, the diagonal value is close to zero, which leads to the substantially larger values associated to the off-diagonal values along column 1. Additionally, we see how the value associated to the homogeneous model along all other columns is significantly smaller than for the other competing models. This indicates its inability of learning from data generated from other models, and therefore receives an RE value close to 0. Consequently, if one only had the RE confusion matrix to compare with, one would reach the same conclusion as before: that the homogeneous model is most different from the other models.

On the other hand, the geostatistical model presents the normalized, off-diagonal RE scores farthest from 1 (after the homogeneous model), when compared to all other models. This, again, alludes to its differences in flexibility and ability to learn from data (which is greater than that of the other, simpler models). Out of all the models, the 9-zoned model can be deemed the most similar to the geostatistical, given that the former obtains the normalized value closest to 1 when the latter generates the data and vice-versa. This explains why it also obtains the second largest RE in the BMS analysis, given that it learns from the data in a similar way as the geostatistical model.

When comparing the two 5-zoned models to each other, we can observe that they do not receive the same score when the flow-and-transport model generates the data. This shows how adding the transport processes and concentration data allows for a larger gain in information. This also explains why the 9-zoned model presents a normalized value closer to 1 when compared to the 5-zoned transport model, than does the 5-zoned flow model. Both the 5 and 9-zoned transport models consider the same processes, and thus allow for a similar, larger gain in information.

5.3.3. IE: Posterior-Based Comparison

The IE-based MCM is shown in Figure 8d. Recall that the results represent the remaining uncertainty in the posterior parameter distribution. Figure B1d, with the non-normalized values, shows that there is little to no variability in the score for each model, independent of which model is generating the data in spite of different RE values, as shown in Figure B1c. This suggests that the IE values can be highly influenced by the prior parameter distribution, represented by CE, which can prevail over the influence of the data, through RE. Therefore, one must pay close attention when comparing models with different prior flexibility, given the strong influence from the prior distribution in the result.

The two 5-zoned models in Figure 8d present little differences between their IE scores. Both models have a similar prior uncertainty (CE), given that they have the same number of uncertain parameters, with the same distributions. Therefore, the results provide information on the direct effect of RE. The slightly greater off-diagonal value (1.1) when the transport model generates the data is, in this case, due to the greater RE value assigned to the transport model. This means that there are small differences between the models' posterior parameter uncertainty. The results support the previous statement that IE should be compared alongside the other scores, especially RE, to determine how much of the IE value is due to the available data and how much is due to the model's prior flexibility.

6. Summary and Conclusions

In this study, we propose to use information-theoretic scores, namely ELPD, RE and IE, to complement BME for model selection and model justifiably analysis. Employing the connection between Bayesian inference and information theory presented in Oladyshkin and Nowak (2019), we illustrate how ELPD, RE, and IE allow to gain additional insight with regards to (a) posterior model fit (ELPD), (b) information gain in Bayesian updating (RE), and (c) remaining posterior parameter uncertainty (IE). We show how these scores can be interpreted in addition to BME and how they come at little to no additional computing cost, given that the most computationally demanding step involves the multiple (N_{MC}) model evaluations.

We test the proposed methodology on a controlled setup made up of four flow-and-transport 2D-groundwater models, where we compare the models using all four Bayesian and information theoretic scores. Each model considers a different spatial hydraulic conductivity distribution, which results in different model flexibility. The results show how one can reach different conclusions based on the criteria being used and the overall goal of the analysis.

We show how RE can provide a comparison criterion in traditional BMS setups, where all models are calibrated against the same data set as well as models when they use different subsets of the observations. However, its use comes at the cost of solely ranking and/or comparing among models based on how useful the data was to them, that is, how much the parameter uncertainty was reduced through Bayesian updating. As the results show, this can sometimes lead to different decisions than with BME-based model selection. For example, RE can also punish models with an already good prior fit, and therefore don't benefit from Bayesian updating.

IE quantifies the posterior parameter uncertainty after applying Bayesian updating. The results show, however, that IE is strongly influenced by the models' prior distribution, to the extent where priors can have a much larger impact than the model fit to the data. Therefore, IE is useful to complement RE scores, but not as a measure on its own.

We highlight how both BME and ELPD are not appropriate for comparing models that use different subsets of the available calibration data, whether those subsets differ in size or consider different types of data with different measurement errors. This limitation arises from the inherent bias introduced by the normalization factor in the (Gaussian) likelihood function, which strongly depends on data set properties. However, information-based scores such as RE and IE are not subject to this bias as they are considered as information scores for parameter (not data) distributions. This proves useful for multi-fidelity comparisons, where we want to quantify, from a Bayesian perspective, how different model configurations respond to different subsets of the available data or given types of data. We showcase this by including a multi-fidelity comparison between two models with the same $\ln(K)$ distribution but considering different processes: a flow-only versus a flow-and-transport model. Consequently, the models can only be calibrated with different subsets of the data. We compare the models using RE and IE, which provided information regarding how useful the additional transport-related observations where to the model.

In future applications, the benefits or RE and IE for multi-fidelity comparisons can be further explored by comparing models under different configurations, for example, different spatial and/or temporal resolutions or different model dimension. This can force the modeler to average observations (e.g., to meet a given mesh's cover), or exclude points (e.g., due to not calculating directly for a given time step). Additionally, if processes are added gradually to a given model, as in multi-fidelity/multi-scales schemes, the methodology could be used to quantify how information is distributed among models of different fidelities. RE and IE can provide useful Bayesian metrics to quantify the usefulness of the data in each step, or help decide if additional complexity/processes and/or more observation data is needed.

Based on the results, we recommend to complement the traditional BME-based analysis with information-theoretic scores for model selection and comparison purposes. The results show how ELPD, RE, and IE provide additional information regarding the complete updating process involved in the Bayesian framework, and come at no significant additional computational cost. We do not wish to influence our reader's decision as to which criteria is best nor which one should be used, but advocate that considering all four criteria can lead to a better-informed decision.

Appendix A: Analysis of the Effect of Data Sets Properties

To mathematically show the effect that the calibration data set properties, namely data set size and measurement error distribution, have on the different scores, we expand the different terms in Equation 17 for RE. Here, we use NF to group the normalization factor in the likelihood function (Equation 2), such that:

$$NF = (2\pi)^{\frac{-N_o}{2}} |\mathbf{R}|^{-1/2}. \tag{A1}$$

Additionally, the difference between the observed and modeled data is shown in its vectorial form:

$$(\mathbf{y}_k - \mathbf{y}_o) = \boldsymbol{\delta}. \tag{A2}$$

Equation A3 shows the simplification of the ln(BME) term based on Equation 4:

$$\begin{aligned} -\ln(BME) &= -\ln\left(\mathbb{E}_{prior}\left[NF \cdot \exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) \\ &= -\ln(\mathbb{E}_{prior}[NF]) - \ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) \\ &= -\ln(NF) - \ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) \end{aligned} \tag{A3}$$

and Equation A4 shows the simplification of ELPD from Equation 15 into its basic components:

$$\begin{aligned} ELPD &= \mathbb{E}_{post}\left[\ln\left(NF \cdot \exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right)\right] \\ &= \mathbb{E}_{post}[\ln(NF)] + \mathbb{E}_{post}\left[\ln\left(\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right)\right] \\ &= \ln(NF) + \mathbb{E}_{post}\left[-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right] \end{aligned} \tag{A4}$$

As can be seen in Equations A3 and A4, both scores depend on the natural logarithm of the normalization factor (cannot be disregarded), which has a high dependence on the number of data points and measurement error variance.

By combining the final simplified formulations in Equations A3 and A4, one can rewrite the equation for RE, based on Equation 17, as follows:

$$\begin{aligned} RE &= \left(-\ln(NF) - \ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right)\right) \\ &\quad + \left(\ln(NF) + \mathbb{E}_{post}\left[\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) \\ &= -\ln(\mathbf{NF}) - \ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) + \ln(\mathbf{NF}) \\ &\quad + \mathbb{E}_{post}\left[\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right] \\ &= -\ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) + \mathbb{E}_{post}\left[\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]. \end{aligned} \tag{A5}$$

In Equation A5, the dependence on the normalization factor NF from both BME and ELPD is canceled out, since it is a constant for each model M_K. Consequently, RE depends solely on the exponential term of the likelihood function.

Appendix B: Bayesian Model Similarity Analysis Results

Figure B1 shows the resulting model confusion matrices for the averaged ln(BME) (a), ELPD (b), RE (c), and IE (d) within the Bayesian model similarity analysis. We can observe the same tendencies in Figure B1 as with the normalized MCM in Figure 8. The latter, however, allows for a more clear interpretation, and focuses on the off-diagonal values, which is why we prefer it to represent model similarities.

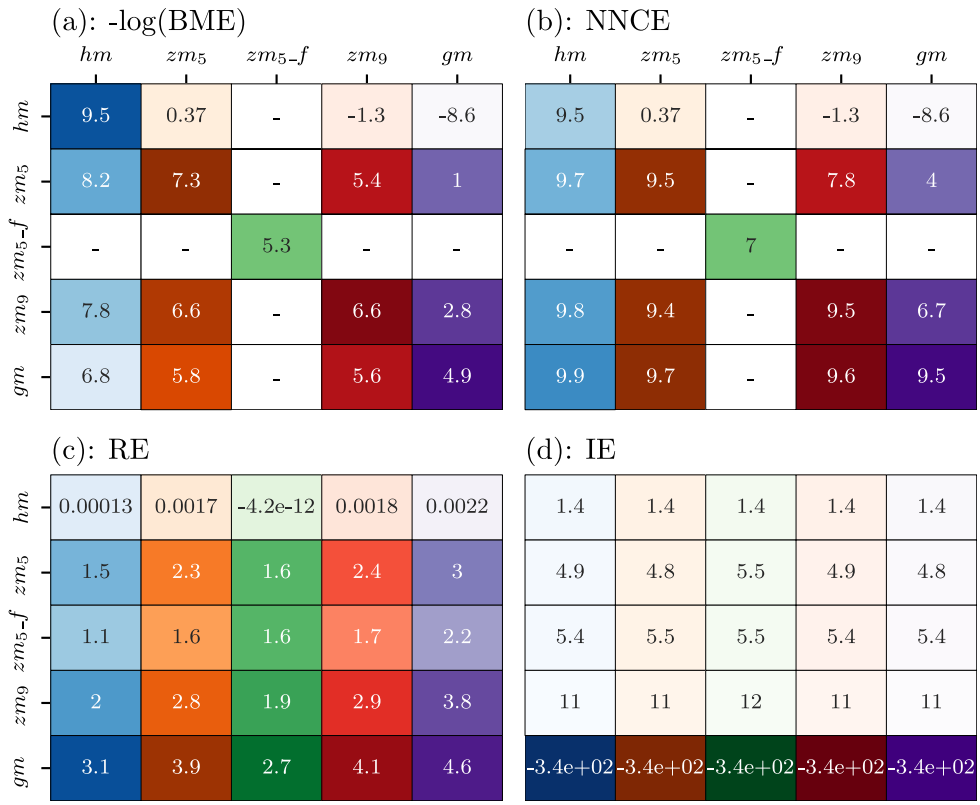

Figure B1. Model confusion matrices for (a) ln(BME), (b) ELPD, (c) relative entropy and (d) information entropy for the Bayesian model similarity analysis. The empty cells in (a) and (b) correspond to the cases where the flow model, which uses a smaller calibration data set, is involved.

Data Availability Statement

The Python implementation of the Bayesian and information-theoretic model selection and similarity analysis can be accessed from the GitHub repository https://github.com/MariaFMoralesOreamuno/Bayesian_Information_theoretic_model_selection.git (Morales Oreamuno, 2021). The files that serve as input for the aforementioned software can be found in https://doi.org/10.5281/zenodo.7086127 (Morales Oreamuno, 2022).

Acknowledgments

We would like to thank Anneli Guthke for her invaluable contributions to this work, including the formulation of the problem and for providing the code for the case study used in this paper. We also thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for supporting this work by funding EXC 2075—390740016 under Germany's Excellence Strategy. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech). We would also like to thank the Bundesgesellschaft für Endlagerung (BGE, Federal Company for Radioactive Waste Disposal) for their support. Open Access funding enabled and organized by Projekt DEAL.

References

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

Aldrich, J. (1997). RA Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, *12*(3), 162–176. https://doi.org/10.1214/ss/1030037906

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

Brunetti, C., Linde, N., & Vrugt, J. A. (2017). Bayesian model selection in hydrogeophysics: Application to conceptual subsurface models of the South Oyster Bacterial Transport Site, Virginia, USA. *Advances in Water Resources*, *102*, 127–141. https://doi.org/10.1016/j.advwatres.2017.02.006

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, *96*(453), 270–281. https://doi.org/10.1198/016214501750332848

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., & Stine, R. A. (2001). The practical implementation of Bayesian model selection. In *Lecture notes-monograph series* (Vol. 38, pp. 65–134). Retrieved from http://www.jstor.org/stable/4356164

Cliff, O. M., Prokopenko, M., & Fitch, R. (2018). Minimising the Kullback–Leibler divergence for model selection in distributed nonlinear systems. *Entropy*, *20*(2), 51. https://doi.org/10.3390/e20020051

Commenges, D. (2015). Information theory and statistics: An overview. https://doi.org/10.48550/ARXIV.1511.00860

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B*, *57*(1), 45–97. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x

Elshall, A. S., & Ye, M. (2019). Making steppingstones out of stumbling blocks: A Bayesian model evidence estimator with application to groundwater transport model selection. *Water*, *11*(8), 1579. https://doi.org/10.3390/w11081579

Enemark, T., Peeters, L. J., Mallants, D., & Batelaan, O. (2019). Hydrogeological conceptual model building and testing: A review. *Journal of Hydrology*, *569*, 310–329. https://doi.org/10.1016/j.jhydrol.2018.12.007

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B*, *56*(3), 501–514. https://doi.org/10.1111/j.2517-6161.1994.tb01996.x

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429258411

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gong, W., Gupta, H. V., Yang, D., Sricharan, K., & Hero, A. O., III. (2013). Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water Resources Research*, *49*(4), 2253–2273. https://doi.org/10.1002/wrcr.20161

Good, I. (1956). Some terminology and notation in information theory. *Proceedings of the IEE-Part C: Monographs*, *103*(3), 200–204. https://doi.org/10.1049/pi-c.1956.0024

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*(8), W08301. https://doi.org/10.1029/2011WR011044

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hammersley, J. M. (1960). Monte Carlo methods for solving multivariable problems. *Annals of the New York Academy of Sciences*, *86*(3), 844–874. https://doi.org/10.1111/j.1749-6632.1960.tb42846.x

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401. https://doi.org/10.1214/ss/1009212519

Höge, M., Guthke, A., & Nowak, W. (2019). The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, *572*, 96–107. https://doi.org/10.1016/j.jhydrol.2019.01.072

James, A. L., & Oldenburg, C. M. (1997). Linear and Monte Carlo uncertainty analysis for subsurface contaminant transport simulation. *Water Resources Research*, *33*(11), 2495–2508. https://doi.org/10.1029/97WR01925

Jung, Y. (2018). Multiple predicting K-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, *30*(1), 197–215. https://doi.org/10.1080/10485252.2017.1404598

Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, *4*(2), 99–104. https://doi.org/10.1109/TPAMI.1982.4767213

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kolmogorov, A. N., & Bharucha-Reid, A. T. (2018). *Foundations of the theory of probability* (Second English edition). Courier Dover Publications.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Lindholm, A., Wahlström, N., Lindsten, F., & Schön, T. B. (2022). *Machine learning: A first course for engineers and scientists*. Cambridge University Press.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005. https://doi.org/10.1214/aoms/1177728069

Liu, C., & Liu, Q. (2012). Marginal likelihood calculation for the Gelfand–Dey and Chib methods. *Economics Letters*, *115*(2), 200–203. https://doi.org/10.1016/j.econlet.2011.12.034

Marsh, C. (2013). *Introduction to continuous entropy*. (p. 1034). Department of Computer Science. Retrieved from https://www.crmarsh.com/pdf/Charles_Marsh_Continuous_Entropy.pdf

Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. *Water Resources Research*, *41*(10), W10422. https://doi.org/10.1029/2004WR003719

Mohammadi, F., Kopmann, R., Guthke, A., Oladyshkin, S., & Nowak, W. (2018). Bayesian selection of hydro-morphodynamic models under computational time constraints. *Advances in Water Resources*, *117*, 53–64. https://doi.org/10.1016/j.advwatres.2018.05.007

Morales Oreamuno, M. F. (2021). Bayesian_information_theoretic_model_selection [Software]. Retrieved from https://github.com/MariaFMoralesOreamuno/Bayesian_Information_theoretic_model_selection.git

Morales Oreamuno, M. F. (2022). Input data for Bayesian and information theoretic model selection and similarity analysis [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7086127

Mouris, K., Acuna Espinoza, E., Schwindt, S., Mohammadi, F., Haun, S., Wieprecht, S., & Oladyshkin, S. (2023). Stability criteria for Bayesian calibration of reservoir sedimentation models. *Modeling Earth Systems and Environment*. https://doi.org/10.1007/s40808-023-01712-7

Murari, A., Peluso, E., Cianfrani, F., Gaudio, P., & Lungaroni, M. (2019). On the use of entropy to improve model selection criteria. *Entropy*, *21*(4), 394. https://doi.org/10.3390/e21040394

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

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

Neuman, S., & Wierenga, P. J. (2003). *A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites*. (Technical Report No. NUREG/CR-6805). U.S.Nuclear Regulatory Commission.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B*, *56*(1), 3–26. https://doi.org/10.1111/j.2517-6161.1994.tb01956.x

Nicenboim, B., Schad, D., & Vasishth, S. (2021). An introduction to Bayesian data analysis for cognitive science. In *Under contract with Chapman and Hall/CRC statistics in the social and behavioral sciences series*. Retrieved from https://vasishth.github.io/bayescogsci/book/

Nowak, W., & Cirpka, O. A. (2006). Geostatistical inference of hydraulic conductivity and dispersivities from hydraulic heads and tracer data. *Water Resources Research*, *42*(8), W08416. https://doi.org/10.1029/2005WR004832

Nowak, W., & Guthke, A. (2016). Entropy-based experimental design for optimal model discrimination in the geosciences. *Entropy*, *18*(11), 409. https://doi.org/10.3390/e18110409

Oladyshkin, S., Mohammadi, F., Kroeker, I., & Nowak, W. (2020). Bayesian^3 active learning for the Gaussian process emulator using information theory. *Entropy*, *22*(8), 890. https://doi.org/10.3390/e22080890

Oladyshkin, S., & Nowak, W. (2019). The connection between Bayesian inference and information theory for model selection, information gain and experimental design. *Entropy*, *21*(11), 1081. https://doi.org/10.3390/e21111081

Press, S. J. (2009). *Subjective and objective Bayesian statistics: Principles, models, and applications* (Vol. 590). John Wiley & Sons. https://doi.org/10.1002/9780470317105

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

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process—A framework and guidance. *Environmental Modelling & Software*, *22*(11), 1543–1556. https://doi.org/10.1016/j.envsoft.2007.02.004

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

Reuschen, S., Nowak, W., & Guthke, A. (2021). The four ways to consider measurement noise in Bayesian model selection—And which one to choose. *Water Resources Research*, *57*(11), e2021WR030391. https://doi.org/10.1029/2021WR030391

Rojas, R., Feyen, L., Batelaan, O., & Dassargues, A. (2010). On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling. *Water Resources Research*, *46*(8), W08520. https://doi.org/10.1029/2009WR008822

Rojas, R., Feyen, L., & Dassargues, A. (2008). Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research*, *44*(12), 416–435. https://doi.org/10.1029/2008WR006908

Santamaría-Bonfil, G., Fernández, N., & Gershenson, C. (2016). Measuring the complexity of continuous distributions. *Entropy*, *18*(3), 72. https://doi.org/10.3390/e18030072

Schäfer Rodrigues Silva, A., Guthke, A., Höge, M., Cirpka, O. A., & Nowak, W. (2020). Strategies for simplifying reactive transport models: A Bayesian model comparison. *Water Resources Research*, *56*(11), e2020WR028100. https://doi.org/10.1029/2020WR028100

Scheurer, S., Schäfer Rodrigues Silva, A., Mohammadi, F., Hommel, J., Oladyshkin, S., Flemisch, B., & Nowak, W. (2021). Surrogate-based Bayesian comparison of computationally expensive models: Application to microbially induced calcite precipitation. *Computational Geosciences*, *25*(6), 1899–1917. https://doi.org/10.1007/s10596-021-10076-9

Schöniger, A. (2010). *Parameter estimation by ensemble Kalman filters with transformed data*. Diplomarbeit, Universität Stuttgart.

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*, 96–110. https://doi.org/10.1016/j.jhydrol.2015.07.047

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. https://doi.org/10.1002/2015WR016918

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. https://doi.org/10.1002/2014WR016062

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Ill. Press Urbana I.

Shore, J., & Johnson, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, *26*(1), 26–37. https://doi.org/10.1109/TIT.1980.1056144

Smith, A., & Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician*, *46*(2), 84–88. https://doi.org/10.1080/00031305.1992.10475856

Smith, J., & Smith, P. (2007). *Environmental modelling: An introduction*. Oxford University Press.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 198–192. https://doi.org/10.1016/j.aci.2018.08.003

Vecer, J. (2019). Dynamic scoring: Probabilistic model selection based on utility maximization. *Entropy*, *21*(1), 36. https://doi.org/10.3390/e21010036

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. https://doi.org/10.1214/12-SS102

Wainwright, J., & Mulligan, M. (2013). *Environmental modelling: Finding simplicity in complexity*. John Wiley & Sons.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594. https://doi.org/10.48550/arXiv.1004.2316

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 Resources Research*, *51*(4), 2825–2846. https://doi.org/10.1002/2014WR016292

Ye, M., Neuman, S. P., & Meyer, P. D. (2004). Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research*, *40*(5), W05113. https://doi.org/10.1029/2003WR002557

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician*, *42*(4), 278–280. https://doi.org/10.1080/00031305.1988.10475585