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[2]

[1]University of Stuttgart
[2]Universität Stuttgart

November 24, 2022

Abstract

Bayesian model selection (BMS) and Bayesian model justifiability analysis (BMJ) provide a statistically rigorous framework to compare competing conceptual models through the use of Bayesian model evidence (BME). However, BME-based analysis has two main limitations: (1) it's powerless when comparing models with different data set sizes and/or types of data and (2) doesn't allow to judge a model's performance based on its posterior predictive capabilities. Thus, traditional BME-based approaches ignore useful data or models due to issue (1) or disregards Bayesian updating because of issue (2). To address these limitations, we advocate to include additional information-theoretic scores into BMS and BMJ analysis: 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 indicate the information flow in BMS and BMJ analysis. We show how to compute and interpret these scores alongside BME, and apply it in a model selection and similarity analysis framework. We test the methodology on a controlled 2D groundwater setup considering five competing conceptual models accompanied with different data sets. The results show how the information-theoretic scores complement BME by providing a more complete picture concerning the Bayesian updating process. Additionally, we present how both RE and IE can be used to objectively compare models that feature different data sets. Overall, the introduced Bayesian information-theoretic framework helps to avoid any potential loss of information and leads to an informed decision for model selection and similarity.

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], **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 BME is not suited to compare models with different data set sizes, while relative entropy can overcome these limitations
- We provide a BMS methodology based on Bayesian and information-theoretic scores, including approaches that help to interpret the results

Corresponding author: Maria Fernanda Morales Oreamuno, maria.morales@iws.uni-stuttgart.de

Abstract

Bayesian model selection (BMS) and Bayesian model justifiability analysis (BMJ) provide a statistically rigorous framework to compare competing conceptual models through the use of Bayesian model evidence (BME). However, BME-based analysis has two main limitations: (1) it's powerless when comparing models with different data set sizes and/or types of data and (2) doesn't allow to judge a model's performance based on its posterior predictive capabilities. Thus, traditional BME-based approaches ignore useful data or models due to issue (1) or disregards Bayesian updating because of issue (2). To address these limitations, we advocate to include additional information-theoretic scores into BMS and BMJ analysis: 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 indicate the information flow in BMS and BMJ analysis. We show how to compute and interpret these scores alongside BME, and apply it in a model selection and similarity analysis framework. We test the methodology on a controlled 2D groundwater setup considering five competing conceptual models accompanied with different data sets. The results show how the information-theoretic scores complement BME by providing a more complete picture concerning the Bayesian updating process. Additionally, we present how both RE and IE can be used to objectively compare models that feature different data sets. Overall, the introduced Bayesian information-theoretic framework helps to avoid any potential loss of information and leads to an informed decision for model selection and similarity.

1 Introduction

Environmental modelling 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 modelling, 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 modelling, including parameter, model input and measurement uncertainty (Refsgaard et al., 2007). Additionally, there is uncertainty associated with the model itself, referred to as conceptual uncertainty, which has been proven

to be a main source of uncertainty (Bredehoeft, 2005; Neuman, 2003; Rojas et al., 2008; Gupta et al., 2012).

Due to incomplete knowledge on the real system, there is not one 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 values, 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 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 (Neuman et al., 2003; Bredehoeft, 2005; Refsgaard et al., 2006). Here, a group of competing conceptual models are either generated or selected, and then tested against some acceptance criteria regarding, e.g, 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 (MSE), Nash-Sutcliffe efficiency (Nash & Sutcliffe, 1970) and cross validation methods (Stone, 1974; Jung, 2018) as model comparison criteria (Gupta et al., 2009). Nevertheless, these do not allow to account for parameter uncertainty, unlike a Bayesian approach.

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 (Schöniger, Illman, et al., 2015). 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 (Raftery, 1995;

Draper, 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 modelling, Wöhling et al. (2015) for crop modelling, Marshall et al. (2005) for hydrological models, Brunetti et al. (2017) in hydrogeophysical modelling 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 to biochemical processes in the subsurface, respectively.

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 said 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 (BIC) (Schwarz, 1978) and the Kayshap information criterion (KIC) (Kashyap, 1982), to name a few. These, however, require assumptions which can lead to biased results (Schöniger et al., 2014). The Monte Carlo sampling technique (Hammersley, 1960) provides a bias-free framework to approximate BME, given that it allows to sample from the entire prior parameter space. 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 analysis does provide a statistically rigorous methodology for considering uncertainties, it does present some limitations, which also extend to the BMJ methodology. The first limitation of the BME-based frameworks consists in BME being powerless when comparing models that can only work with different data sets sizes and/or types of data. This can be the case for models with different space discretization, which use a different subset of the available data, or models with different levels of complexity, e.g. solely flow or flow-transport models, that can reproduce different types of data. BME, through the likelihood function, is dependent on data set properties, such as data set size and measurement error. When comparing models with different data sets, this influence can be independent on model fit. Therefore comparing models with different data sets can lead to biased results. This implies that any extra, potentially informative data associated to a subset of models cannot be used, and would therefore be wasted. Several studies have addressed the impact of measurement error, data type and data set size on model rankings (see Rojas et al., 2010; Schöniger, Wöhling, & Nowak, 2015; Wöhling et al., 2015), and have shown that one can obtain significantly different BME weights depending on the number and/or type of data used for the comparison. These studies, however, have focused on model sets within the traditional BMS framework (comparing all models against the same data) and have not, to the best of our knowledge, addressed the subject of comparing models based on different data sets and the direct effect of the likelihood function.

Indeed, BME, through the likelihood function, depends on the data set properties, including measurement quality (error) and data set size. Schöniger, Wöhling, and Nowak (2015) and Wöhling et al. (2015) mention that one could get significantly different model selection scores (or model weights) depending on the data set chosen for calibration or testing. Consequently, the methodology is limited to comparing models with the exact same calibration/testing data set to avoid bias in the results.

The second limitation of the existing BME-based approaches is that BME does not allow to judge the performance of a model based on its posterior predictive capabilities. Basically, in the Bayesian theory (see Kolmogorov & Bharucha-Reid, 2018; Gelman et al., 1995), BME is considered as a normalization factor that can be obtained via the in-

tegration of the likelihood over the prior parameter space. Therefore, BME values contain only partial information required for the Bayesian updating of a model via the observation data. Additionally, they 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.

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 the latter is linked to maximum entropy quantification and is efficient in terms of information content (Zellner, 1988). Information theory scores include the expected-log posterior likelihood (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 (Gelman et al., 2014; Murari et al., 2019; Cliff et al., 2018; Vecer, 2019) and optimal experimental design (Nowak & Guthke, 2016; Lindley, 1956).

Many applications use approximations of entropy, such as the Akaike information criteria (AIC) (Akaike, 1974), 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, however, require the use of assumptions, which can cause bias in the results (Oladyshkin & Nowak, 2019). To overcome this, Oladyshkin and Nowak (2019) present a connection between Bayesian inference and information theory and propose methods to compute BME in combination with ELPD, RE and IE, to measure information content in Bayesian updating and for model selection purposes.

Based on the methods proposed in this study, some of these information-theoretic scores remain meaningful when comparing models with different data sets, given that they tend to be less dependent on data set properties. Moreover, as suggested in Oladyshkin and Nowak (2019), the relation between Bayesian inference and computing certain information-theoretic scores avoids any additional assumptions and skips any multidimensional integration or density estimation, which is why we have chosen to follow said methodology. Overall, the Bayesian information-theoretic scores allow to obtain information on

the updating process within the Bayesian inference framework, which is ignored in traditional BME-based BMS and BMJ analysis. The approach proposed in (Oladyshkin & Nowak, 2019) has been applied in active learning techniques for surrogate model generation (Oladyshkin et al., 2020), but not, to the authors' knowledge, for model selection or similarity analysis as in the context of the current paper.

The current paper proposes to complement the traditional BME-based BMS methodology with information-theoretic scores. The goal is to overcome the two aforementioned limitations surrounding BME. We focus mainly on the addition of ELPD as a measure of information between the likelihood and the posterior (posterior model fit), RE between the prior and the posterior (updatability conditioned on the data) and IE of the posterior for model selection and comparison purposes. To avoid additional assumptions due to the novelty of the methodology, we make use of the prior-based Monte Carlo sampling approach (Hammersley, 1960; Gelman et al., 1995). Additionally, and building on the work by Schöniger, Illman, et al. (2015), we seek to further complement the BMS procedure with a model similarity analysis, using model confusion matrices based on BME and the different information-theoretic scores, to determine in which step of the Bayesian updating process do the models present similarities, or differences.

We apply and test the methodology on a synthetic groundwater model setup, made up of five competing models and based on the setup in Schöniger, Illman, et al. (2015). Here, four flow-transport models and one flow-only model, with different spatial hydraulic conductivity distribution, are compared against each other (model similarity analysis) and against a set of synthetically generated data (BMS). This setup will allow to test our proposed methodology on environmental models with different complexity, represented by their priors, as well as with different data sets. With this study and its application case, we seek to 1) present the behavior of the information-theoretic scores within the BMS and model similarity frameworks, and how they can be interpreted to complement BME; 2) determine which scores can be used to select and compare between models with different data sets, and the limitations associated to them.

The remainder of the paper is organized as follows: in Section 2 we present an overview of traditional BME 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. This distribution can be the result from previous experience or field measurements (Moore & Doherty, 2005). The current beliefs are then updated based on how well the model can reproduce historically observed data to obtain a posterior distribution (Raftery, 1995), which should be more, or just as, informative as the prior (Oladyshkin & Nowak, 2019). Bayes' theorem can therefore 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, $p(\mathbf{y}_o|\boldsymbol{\omega}_k, M_k)$ 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 Bayesian model evidence (BME).

The likelihood function (see Aldrich, 1997) serves as the connection between the prior and posterior distributions by incorporating the observed data (Press, 2009). Specifically, the term $p(\mathbf{y}_o|\boldsymbol{\omega}_k, M_k)$ in equation (1) states how likely it is that a given model M_k, with parameter set $\boldsymbol{\omega}_k$, can fit the observed data set $\mathbf{y}_o$ within the tolerance implied by the data's measurement error. 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 x 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 error, 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:

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

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

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

where the BME value is expressed as an integral over the total parameter space $\boldsymbol{\omega}_k$ 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 Schöniger et al., 2014; Oladyshkin & Nowak, 2019). However, equation (3) is often employed using the prior-based brute Monte Carlo (MC) sampling (Hammersley, 1960), yielding to the following estimate:

$$BME_k \approx \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} p(\mathbf{y}_o|M_k, \omega_i). \tag{5}$$

It is well-known that MC sampling in equation (5) 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 (see details in Schöniger, Illman, et al., 2015; Oladyshkin & Nowak, 2019). Therefore, in the current paper, we follow the MC sampling strategy to avoid additional assumptions and biased results, given the novelty of the framework we are suggesting.

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):

$$W(M_k|\mathbf{y}_o) = \frac{p(\mathbf{y}_o|M_k)W(M_k)}{\sum_{i=1}^{N_M} p(\mathbf{y}_o|M_i)W(M_i)}, \tag{6}$$

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 (6) is a normalizing factor, such that the sum of all model posterior probabilities is equal to 1. This denominator is the same across all models. Therefore, the only term which has an effect on the posterior model weight is $p(\mathbf{y}_o|M_k)$, which is the BME for model M_k, that 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, where the latter is represented by the prior distribution. Importantly, BME values are valid only for the current state of knowledge, and are dependent on the data and 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) could change.

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 modelling 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 $l = 1 : N_d$, 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 ($BME_{k,l}$) are then averaged over all N_d realizations, to obtain an averaged $BME_{k,j}$ value, as summarized by the following equation:

$$\overline{BME}_{k,j} = \frac{1}{N_d} \sum_{l=1}^{N_d} BME_{k,l} \,, \tag{7}$$

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 model confusion matrix (MCM) (Schöniger, Illman, et al., 2015). The MCM has the size N_M x 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, e.g. classification problems (see Lindholm et al., 2022).

Figure 1 shows a schematic illustration of generating the MCM. Following the order set by equation (7), 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 $i = 1...N_{MC}$, 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 $BME_{k,l}$. 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 $BME_{k,l}$ values are averaged to obtain the MCM entries, represented by the yellow area in Figure 1.

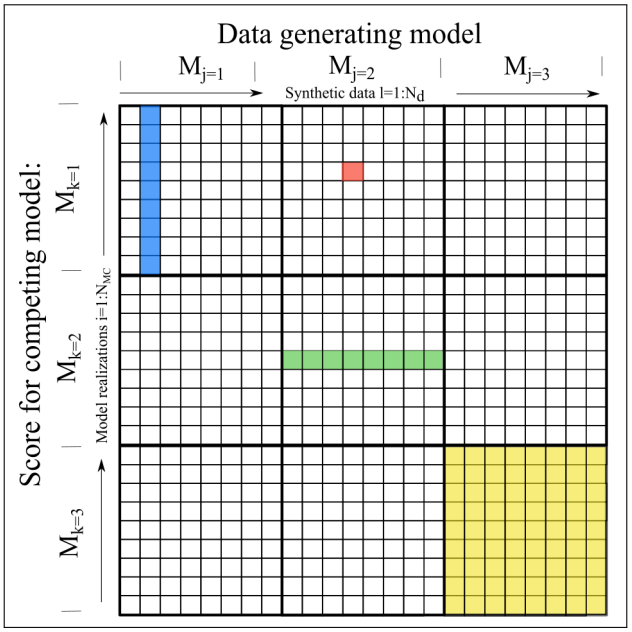

Figure 1. Schematic illustration how to construct a model confusion matrix. Red box: likelihood of a single realization drawn from model $M_{k=1}$, given a single realization drawn from model $M_{j=2}$. Blue boxes: average likelihood (BME) of model $M_{k=1}$ given a single realization of model $M_{j=2}$. 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.

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 calibration data set properties, 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 applying equation (6) to calculate $W(M_k|\mathbf{y}_o)$. In this case, the effect of data set size and measurement error is concentrated inside the exponential term, in the values and size of $\mathbf{R}$, where it is combined with 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. Thus, their effect on the normalization factor must be taken into account and will directly affect the BME value, independent on model fit. 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 values, and thus BME. In the second term, $|\mathbf{R}|^{-1/2}$, the value of the determinant of $\mathbf{R}$ depends on both data set size and the magnitude of the measurement error. Consequently, BME becomes powerless when comparing models with different data sets, since the models would not be under equal conditions and thus BME would lead to biased results. We will further explore this scenario using a groundwater case study, in which we compare models with different data sets. We will use this example to expose the problems with BME in these cases, and how we can use information-theoretic scores to potentially 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 modelling, especially with regard to the spatially variable 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), where the problem of choosing a

spatial parametrization for hydraulic conductivity (heterogeneity) given a set of possible models is addressed. Additionally, one could also be confronted with the problem of choosing which processes to consider as relevant, for example whether we use a flow-only or a flow-and-transport model to represent our system. A reason to do so is that additional transport-related data can be highly informative on details in hydraulic conductivity fields, but only if a corresponding upgrade of models to include transport (and transport-related parameters) is done (Nowak & Cirpka, 2006). If one considers both flow-and-transport and flow-only models, the competing models would therefore depend on different data sets. Consequently, it poses a challenge for the traditional BME-based BMS approach described in the previous section. To overcome this challenge, we propose to test this scenario, considering models with different data sets, with the Bayesian and information-theoretic methodology.

We set up a synthetic groundwater model comparison, where the challenge of modelling subsurface heterogeneity is examined by comparing four models with different spatial distributions of hydraulic conductivity (K). This setup is loosely based on the one presented in Schöniger, Illman, et al. (2015). We build on this work by additionally considering both flow and transport models, which depend on different data sets. Consequently, the following five competing 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).

We will compare the models against 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 synthetically true K distribution.

Through this application, we seek to demonstrate the behavior of the additional Bayesian information-theoretic scores for models with different conceptual representations (prior flexibility) and with different data sets. We also plan to address how the models with different data sets could be assessed against each other using MCM. To do so,

we will first summarize the simulation setup as well as the competing models in the following sections, 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 (FEM) code, based on the program used in Schöniger (2010). The program solves the steady state, 2D groundwater transport equations for a 50 m x 50 m confined aquifer, discretized every 1 m. A Dirichlet boundary condition of 1 m 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. 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).

Table 1. Boundary conditions and constant aquifer and transport parameters

Parameter	Value
Domain size	$[50\ m, 50\ m]$
Grid size	$[1\ m, 1\ m]$
West BC*	1 m
East BC*	0 m
North BC*	$0\ m/s$
South BC*	$0\ m/s$
Porosity	0.35
Longitudinal dispersivity	$2.5\ m$
Transverse dispersivity	$0.5\ m$
Diffusion coefficient	$1x10^{-9}\ m/s$

*BC = Boundary condition.

We consider four different hydraulic conductivity (K) models to generate five competing groundwater models, following the logic presented in Schöniger, Illman, et al. (2015). The homogeneous model represents the simplest model, since it consists of a single K value assigned to all cells in the grid. The controlled nature of the experiment allows for two informed, zoned models, one divided into five independent K zones, and one divided into nine zones, with the latter therefore being more flexible. For these three models, we assume that the ln(K) values follow a normal distribution with a mean of $\ln(1x10^{-5})$ 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 $\ln(1x10^{-5})$, a variance of 1 and correlation length of [10 m.,10 m.]. This results in 2500 uncertain parameters, which are all dependent on each other. A summary of the different ln(K) parametrization models can be seen in Table 2.

For this test case, the synthetically true ln(K) distribution was generated from a realization of the geostatistical model, which can be observed in Figure 2. The synthetic setup and the synthetic observation data generated from it will be discussed further in Section 3.2. The informed zone classification for both the 5-zoned and the 9-zoned models were based on this ln(K) distribution to simulate a prior knowledge of the real ln(K) field. Both zone classifications can be seen in Figure 3.

Table 2. Summary of hydraulic conductivity parametrization models

Model	Number of parameters	Parameters' distribution
Homogeneous (hm)	1	$\mathcal{N}$ [ln($1x10^{-5}$), 1]
5-zoned (zm_5)	5	$\mathcal{N}$ [ln($1x10^{-5}$), 1]
9-zoned (zm_9)	9	$\mathcal{N}$ [ln($1x10^{-5}$), 1]
geostatistical (gm)	2500	$\mathcal{N}$ [ln($1x10^{-5}$), $\mathbf{\Sigma}^*$]

$^*\Sigma$ = Exponential covariance function, with correlation length $(x, y) = [10\ m, 10\ m]$.

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 head (h) and concentration (c_o) mea-

surements. 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.

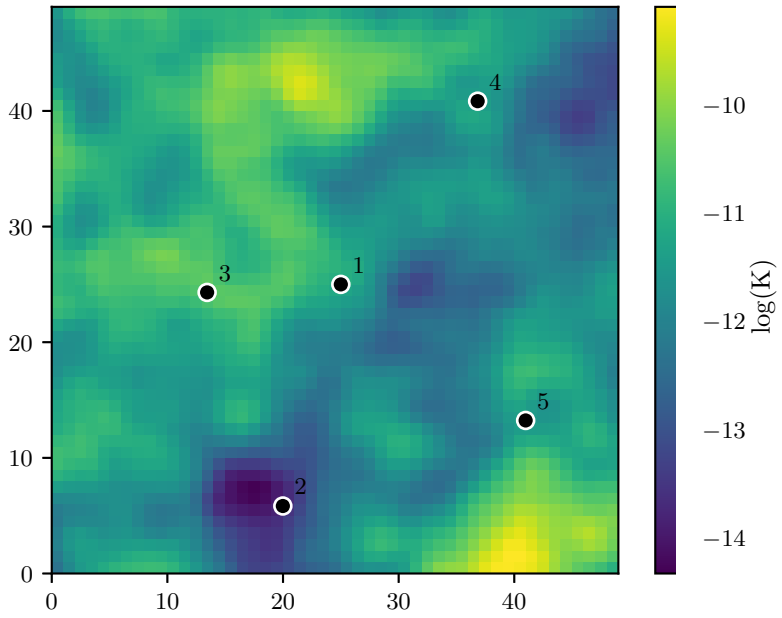

Figure 2. True $\ln(K)$ spatial distribution, synthetically generated through the geostatistical model. The black dots correspond to the location of the measurement points.

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, which 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 $h_{error} = 0.06m$ and $c_{error} = 0.06 + 0.2c_o$,

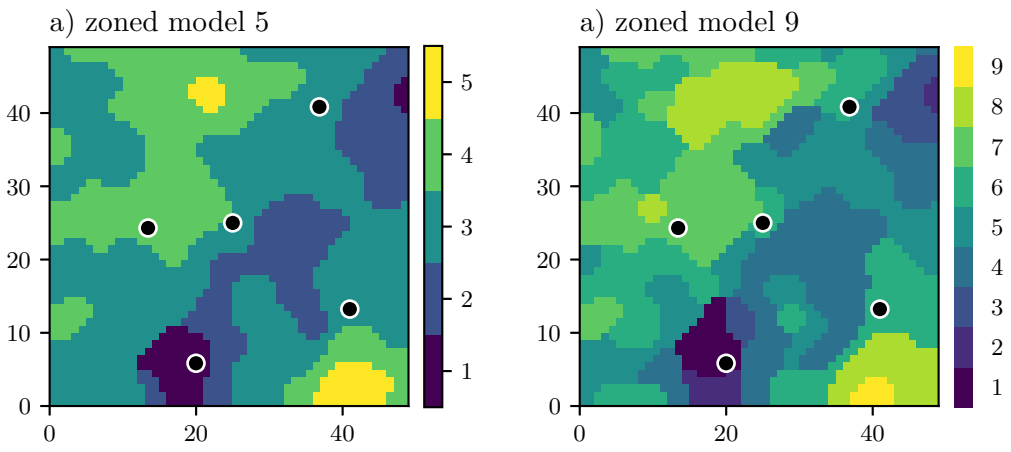

Figure 3. Zone classification for a) 5-zoned model and b) 9-zoned model, based on synthetically true $\ln(K)$ distribution.

assuming a relative error for c_o dependent on the measured value. Therefore, the flow and transport models have not only different data set sizes, but also observations with different measurement errors.

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 or uncertainty in data, referred to as information entropy. Originally, information theory was introduced for discrete probabilities (Shannon, 1948) and then expanded to continuous distributions. Differences with regards to discrete and continuous entropy are further 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 information entropy (IE), cross entropy (CE) and relative entropy (RE) within the Bayesian frame-

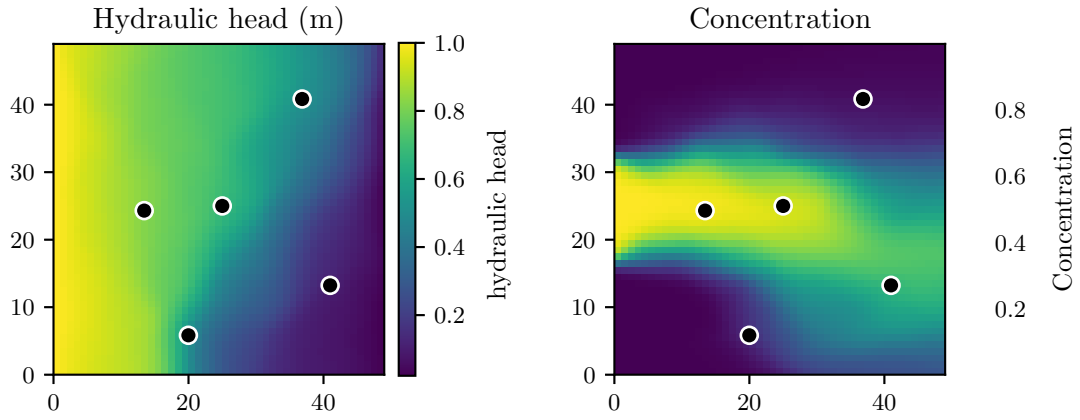

Figure 4. Spatial distribution of hydraulic head (left) and concentration (right) for the true synthetic run, generated with the geostatistical model

work. 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:

$$IE \equiv H[p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o)] = -\int_{\Omega} p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o)\, \ln\left[p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o)\right] d\boldsymbol{\omega}_k, \tag{8}$$

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

$$H[p(x)] = -\int p(x) \cdot \ln[p(x)] dx. \tag{9}$$

Cross entropy (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 pos-

terior $p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o)$ from the prior $p(\boldsymbol{\omega}_k, M_k$ as follows:

$$CE \equiv H[p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o), p(\boldsymbol{\omega}_k, M_k)] = -\int_{\Omega} p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o)\ \ln\left[p(\boldsymbol{\omega}_k, M_k\right] d\boldsymbol{\omega}_k, \tag{10}$$

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

$$H[p(x), q(x)] = -\int p(x) \cdot \ln[q(x)]dx. \tag{11}$$

Similar to the CE in equation (10), 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):

$$NNCE \equiv \hat{H}[p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o), p(\mathbf{y}_o|\boldsymbol{\omega}_k, M_k)] = -\int_{\Omega} p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o) \ln\left[p(\mathbf{y}_o|\boldsymbol{\omega}_k, M_k)\right] d\boldsymbol{\omega}_k. \tag{12}$$

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 (12), 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 (12) represents a posterior-based expectation of the log-likelihood over the entire parameter space:

$$ELPD = -NNCE. \tag{13}$$

ELPD, in its different approximations, has been used to compare and check 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 relative entropy 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:

$$RE \equiv D_{KL}[p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o), p(\boldsymbol{\omega}_k, M_k] = \int_{\Omega} \ln \left[\frac{p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o)}{p(\boldsymbol{\omega}_k, M_k} \right] p(\boldsymbol{\omega}_k, M_k|\mathbf{y}_o) d\boldsymbol{\omega}_k. \tag{14}$$

Using equations (8) and (12), equation (14) can be also rewritten as the difference between the cross entropy (CE) and the information entropy 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:

$$RE = CE - IE. \tag{15}$$

4.2 Computation of Information-Theoretic Scores

Various problems arise when solving equations (8), (12) and (15). 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 multi-dimensional integration.

4.2.1 ELPD

To compute ELPD (and therefore NNCE), we use a brute force Monte Carlo methodology. Given that the posterior parameter and output distributions are usually not known in analytical form, equation (12) can be rewritten as a sample-wise expectation of the posterior (giving equal weights to each posterior sample):

$$ELPD = \mathbb{E}_{post} \left[\ln[p(\mathbf{y}_o|\boldsymbol{\omega}_k, M_k)] \right], \tag{16}$$

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 (16) by:

$$ELPD \approx \frac{1}{N_{post}} \sum_{i=1}^{N_{post}} \ln[\, p(y_o|\boldsymbol{\omega}_i, M_k)\,], \tag{17}$$

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

One can observe a similarity between equation (16) for ELPD and equation (5) for BME; they are both measurements of model fit, with the former being in the posterior and the latter in the prior parameter space. 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 the prior 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) for more details):

$$RE = -\ln[BME] - NNCE, \tag{18}$$

which is also equivalent to:

$$RE = -\ln[BME] + ELPD. \tag{19}$$

Equation (18) indicates that RE can be calculated based on BME and NNCE, both of which can be approximated using MCMC or rejection sampling techniques, 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.

4.2.3 Cross Entropy

The cross entropy between the prior and posterior distributions in equation (10) can be obtained from its definition using the posterior-based expectations (similar to ELPD):

$$CE = -\mathbb{E}_{post}\left[\ln p(\boldsymbol{\omega}_k, M_k)\right] \tag{20}$$

or, numerically, using posterior-based sampling:

$$CE \approx= -\frac{1}{N_{post}} \sum_{i=1}^{N_{post}} \ln[\, p(\boldsymbol{\omega}_k, M_k)\,]. \tag{21}$$

4.2.4 Information Entropy

With knowledge on ELPD, CE and RE, one can calculate IE in the Bayesian context directly from equation (8):

$$IE = CE - RE \tag{22}$$

or

$$IE = \ln\left[BME\right] - ELPD + CE. \tag{23}$$

As previously mentioned, IE is the uncertainty of 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 (22). 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

As ELPD is a likelihood-based score, it is sensitive to both data set size and measurement error. Recall that this is evident from the normalizing factor in equation (2) that cannot be canceled out when comparing models with different data sets. This can be seen in more detail in equation (31) in Appendix A, where the equation for ELPD is decomposed to mathematically see the effect of the normalizing factor. Therefore, ELPD is subject to similar issues as BME, and should only be used to compare models with the same data set.

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 mathematically in equation (32) in Appendix A. When estimating RE, the normalizing factor from the likelihood function, present in both ln(BME) and ELPD, is cancelled 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. 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 Monte Carlo approach, which is the most computationally demanding step) and the rejection sampling processes, 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 of 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:

$$\overline{ELPD}_{k,j} = \frac{1}{N_d} \sum_{l=1}^{N_d} ELPD_{k,l} \,, \tag{24}$$

$$\overline{RE}_{k,j} = \frac{1}{N_d} \sum_{l=1}^{N_d} (-\ln[BME_{k,l}] + ELPD_{k,l}) \,, \tag{25}$$

$$\overline{IE}_{k,j} = \frac{1}{N_d} \sum_{l=1}^{N_d} (-RE_{k,l} - CE_k). \tag{26}$$

The results will comprise of four different MCMs, one for each BMS score. 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:

$$\overline{\ln[BME]}_{k,j} = \frac{1}{N_d} \sum_{l=1}^{N_d} \ln[BME_{k,l}] \,, \tag{27}$$

In contrast to Schöniger, Illman, et al. (2015), we do not calculate Bayesian model weights, since these can only be obtained from BME. Therefore, 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). Consequently, the diagonals in the final MCM will always be equal to 1 and the off-diagonals will indicate how much model

M_k differs, on average, from the self-identification score. 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. This 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).

4.5 Interpretation of MCMs

Based on the description of the different scores, the MCMs for ln(BME) and ELPD can be considered as likelihood-based comparisons. 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). Just like in BMS, we cannot include models with different data sets in the MCMs for BME and ELPD.

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 average, we do not expect the RE values in the diagonal to tend to zero (when the data-generating model is compared against itself). Based on this, two models can be considered similar, from an 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 (22), 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 (5), (17), (19) and (23). For this, we sample $1x10^6$ Monte Carlo realizations from each of the five competing models. For BMS (Section 5.2) we use a sin-

gle 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 (27), (24), (25) and (26) are used to populate the MCMs for ln(BME), ELPD, RE and IE. Here, $N_d = 1000$ Monte Carlo realizations of each possible data-generating model M_j are sampled and then compared to the $N_{MC} = 1x10^6$ Monte Carlo 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.

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). There, we show when the likelihood-based scores (BME and ELPD) can no longer be applied and one must switch to a solely RE and IE-based 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 these results for three different analyses on the behavior of the different scores: 1) comparing the four flow-transport models, using all four scores 2) comparing the five competing models, including the one with a different data set, based on RE and IE and 3) comparing between the two 5-zoned models. Since BME and ELPD are powerless to compare models with different data sets, these 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

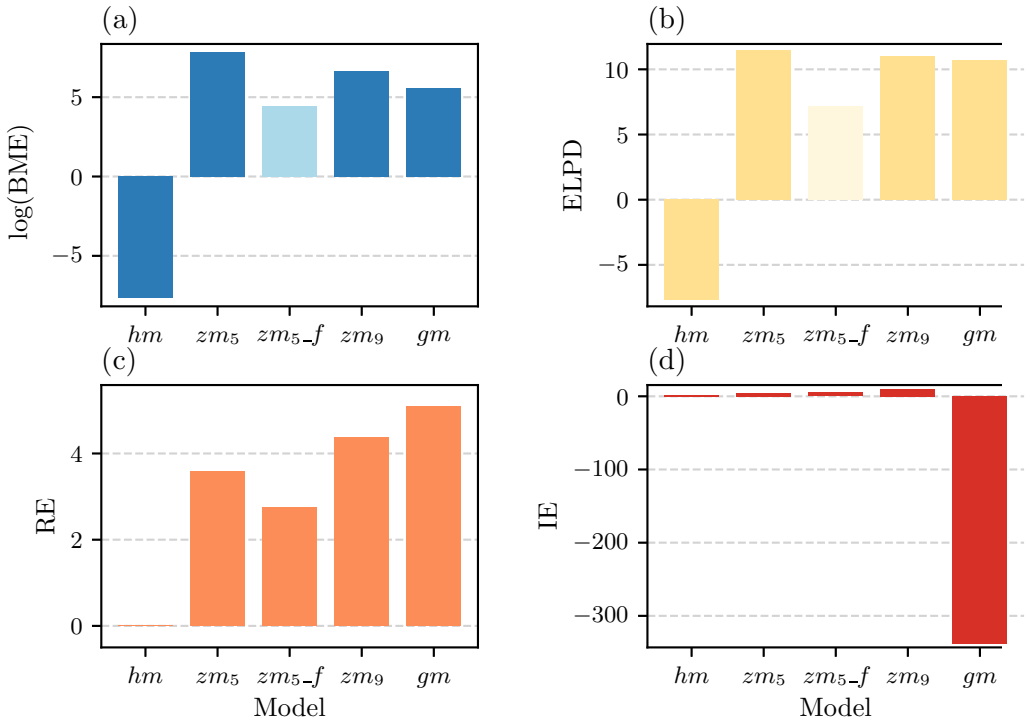

Figure 5. Bayesian and information-theoretic scores for each competing groundwater model in the BMS setup. a) ln(BME), b) expected log-predictive density, c) relative entropy, d) information entropy. The bars with the lighter hues in a) and b) indicate that the corresponding flow model cannot be compared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

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* and would therefore win against models with a lower score. For example, from Figure 5b one would determine that any of the other transport models would be chosen over the homogeneous transport model, given the latter's significantly smaller ELPD (worst overall model fit). The three favored models (i.e. the 5-zoned, 9-zoned and the geostatistical transport model) present similar ELPD scores of 11.46, 11.09 and 10.65,

641 respectively. This means that they have similar posterior predictive capabilities. In other
642 words, all three models have posterior parameter distributions that can similarly pre-
643 dict the observed values. This goes to show how ELPD is less dependent on prior choice
644 when compared to BME. The 5-zoned model, however, still presents the slightly higher
645 ELPD among all four competing models, and would therefore be selected from a pos-
646 terior perspective.

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

659 Even though we have mentioned that models with different data sets should not
660 be compared to each other using BME and ELPD, mathematically it can be done to ob-
661 serve the direct effect of different data sets on said scores. Therefore, if we compare the
662 BME and ELPD values for the two 5-zoned models (with and without transport), we
663 can observe that the transport model, with a larger data set, presents higher scores than
664 the flow model. This can be explained by the relatively small measurement error asso-
665 ciated to the additional concentration data: overall, it is harder for the transport model
666 to reproduce a larger data set. Nevertheless, in this case, the likelihood function rewards
667 the (few) realizations that are able to reproduce all 10 observations within the error thresh-
668 old with a significantly higher likelihood, increasing the expected BME and ELPD val-
669 ues. This indicates how the measurement error and the size of data set can play an im-
670 portant role when calculating BME and ELPD, deeming them biased when comparing
671 models with different calibration data sets.

672 ### *5.2.3 RE-based Selection: Maximization of Relative Information Gain*

673 As opposed to the BME and ELPD-based approaches, RE allows to compare all
674 models, regardless of the data set used by each model. The model with the largest RE
675 represents the one who found the data most useful, meaning it reduced its uncertainty
676 the most by moving from the prior to the posterior. For example, when comparing the
677 two 5-zoned models in Figure 5c, which present different data sets, the transport model
678 obtained a higher RE than the flow-only model. This means that the 5-zoned transport
679 model was able to learn more from the 10 data points than the flow model from only 5
680 observations. Here, the differences in the RE scores are not due to model flexibility, since
681 they both start from the same prior, but due to the availability and informativeness of
682 additional observations available to the transport model. Based on this, one would lean
683 towards the transport model over the flow model, since it allows a greater information
684 gain from prior to posterior while also using the entire available data set.

685 Looking at the overall RE-based analysis in Figure 5c, we notice that the geosta-
686 tistical model presents the highest RE. This indicates that it gained the most informa-
687 tion from the data when updating the prior to posterior, as measured by in RE through
688 the difference of ln(BME) and ELPD (see equation (19)). This goes to show how RE and
689 BME do tend to be inversely related, given that the geostatistical model also obtained
690 the smallest BME out of the top three transport models. We demonstrate that RE tends
691 to favor the models independent from their complexity, but rather when the measure-
692 ment data prove most useful to it. In some situations, the ability of a model to learn from
693 the measurement data could coincide with model flexibility, but the latter should not
694 be seen as a necessary nor sufficient condition. In the current case, the setup was built
695 such that the geostatistical model was able to gain the most information, given that it
696 generated the data. The fact that the 9-zoned model obtained a higher RE and a smaller
697 BME value than the 5-zoned transport model further supports this claim.

698 Additionally, a small RE does not necessarily indicate a bad fit to the measurement
699 data, but can also be caused by an initially good prior fit. This can be seen for the 5-
700 zoned transport model. The smaller RE associated to it can be explained due to an over-
701 all good prior fit to the data (small difference between ln(BME) and ELPD). In other
702 words, it had little to learn from the data given that the prior parameter distribution
703 encompassed the true posterior quite well. This can be seen as a limitation when com-

paring models solely 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. Therefore, we would like to emphasize to the reader that BME, ELPD and RE can be used to complement each other (when possible) and allow to rank and select among models based on different perspectives or goals set by the modeler.

5.2.4 IE-based Selection: Minimization of Posterior Uncertainty

As with RE, IE can also overcome the limitations of BME and ELPD when encountering models with different data sets. Recalling from Section 4.2, IE is the posterior uncertainty associated to the posterior state, and one would therefore select a model with the smallest IE associated to it. 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 significant as that between IE values. We can therefore conclude from equation (23) that the large difference in IE is due to the prior uncertainty through the CE, and not necessarily to a greater gain in information from the data. This goes to show the large influence that the prior parameter distribution has 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.

Furthermore, the reason for the significantly smaller CE of the geostatistical model can be attributed to the curse of dimensionality (Altman & Krzywinski, 2018) (given the 2500 uncertain parameters associated to this model) and the high correlation between parameters. Due to these factors, the space where all parameters are within the allowed prior variance is very small, causing each parameter set to have a high probability density associated to it, which translates to a small entropy. It is worth mentioning that,

736 if the correlation between the parameters were to substantially decrease, the entropy would
737 increase, given that entropy is maximized for increasingly independent parameters.

738 The opposite happens when the parameters are independent, as in the case of the
739 homogeneous and the zoned models: the probability density associated to each realiza-
740 tion decreases with a higher parameter dimension (given parameter independence), and
741 thus the entropy increases. If we omit the geostatistical model for visualization purposes,
742 as displayed in Figure 6, the homogeneous model presents the smallest IE within the re-
743 maining subset. Here, IE indicates that the homogeneous model has a lower posterior
744 parameter uncertainty than the 5 and 9-zoned models. This, however, can be attributed
745 to the prior distribution, more specifically to the number of uncertain parameters, and
746 not to the overall Bayesian updating process. This is supported by the the small BME,
747 ELPD and RE scores associated with the homogeneous model. Additionally, the 9-zoned
748 model got the highest IE score among this subset of models, in spite of it presenting the
749 largest RE among them. It is clear, then, that the IE depends on both the prior param-
750 eter uncertainty and how useful the data is in eliminating the uncertainty specified by
751 this prior. However, the dependence of CE on the number of parameters generates bi-
752 ased results when comparing models with different parameter dimension and should there-
753 fore be avoided in such cases.

754 If we compare both 5-zoned models, which have the same prior assumptions and
755 different data set sizes, we can observe the effect that the data, through RE, has on IE
756 directly. In this case, the transport model presents the smaller IE of the two, since it learned
757 more from the observation data (higher RE). This resulted in a greater reduction in the
758 (initially identical) prior uncertainty (CE). This complements the conclusions reached
759 using the RE score between these two models, but from a posterior parameter uncertainty
760 perspective.

761 ## 5.3 Bayesian Model Similarity Analysis

762 ### *5.3.1 BME and ELPD: Likelihood-based Comparison*

763 To analyze the similarities, or differences, between the transport models in their
764 prior states, one could limit oneself to the original BMJ analysis based on BME-weights,
765 which is presented in Figure 7. Here we can see that the 3rd row (where all transport
766 models M_j are compared to the flow-only model) has to be left empty, since the flow model

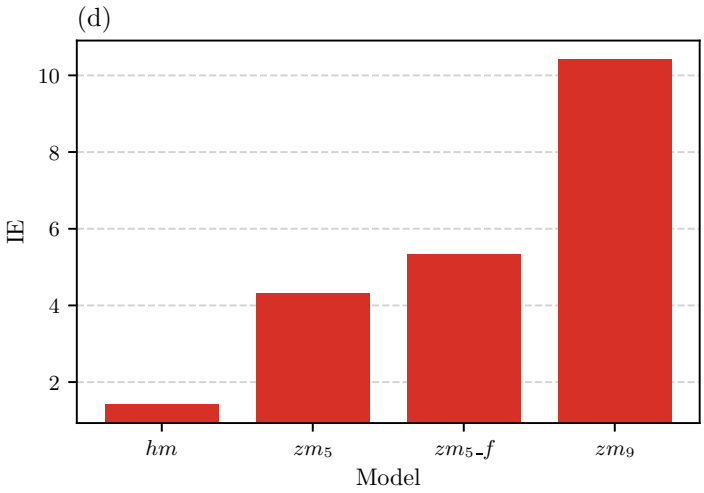

Figure 6. Information entropy scores for all competing groundwater models within the BMS setup, excluding the geostatistical model

has a smaller data set. Additionally, the third column, where the flow model generated the data, is also empty. The reason is that, even though these values can be calculated, they cannot be used for comparison against all other BME weights.

From the results in Figure 7 we can observe that both the homogeneous and the geostatistical model receive high diagonal values, indicating their capacity 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 two 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.

The extended model similarity analysis to determine model similarities, 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 MCM based on ln(BME) (8a) and ELPD (8b) show empty rows and columns where the 5-zoned flow model is compared to the trans-

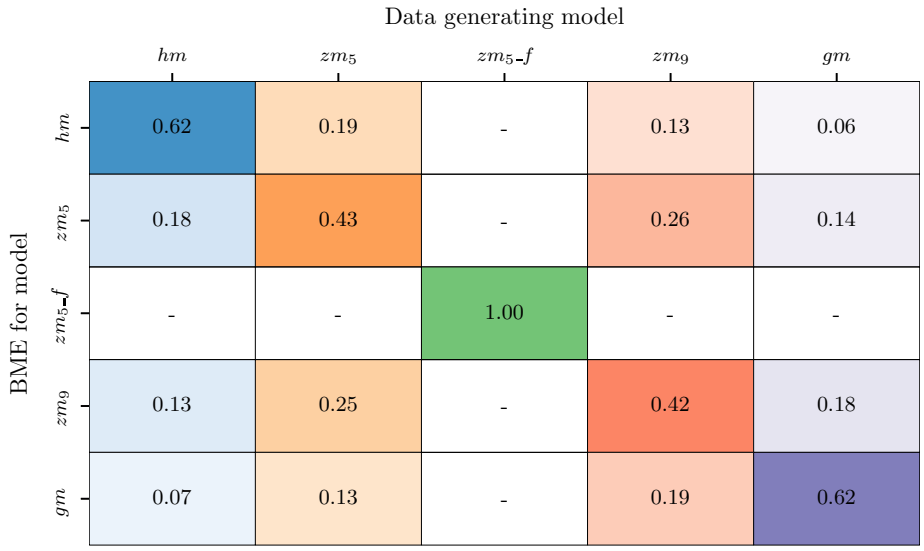

Figure 7. Model confusion matrix based on BME weights. Columns correspond to data-generating models M_j and rows to competing models M_k.

port models. The non-normalized version of the MCMs can be seen in Figure 9 in Appendix B.

As with the BME-weight-based MCM, the ln(BME)-based MCM in Figure 8a compares model outputs from a prior predictive space 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. This can be interpreted as them presenting 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 compared models from a posterior predictive capabilities perspective. This means, how likely model M_k's posterior distribution is of reproducing outputs generated by M_j. The results in Figure 8b

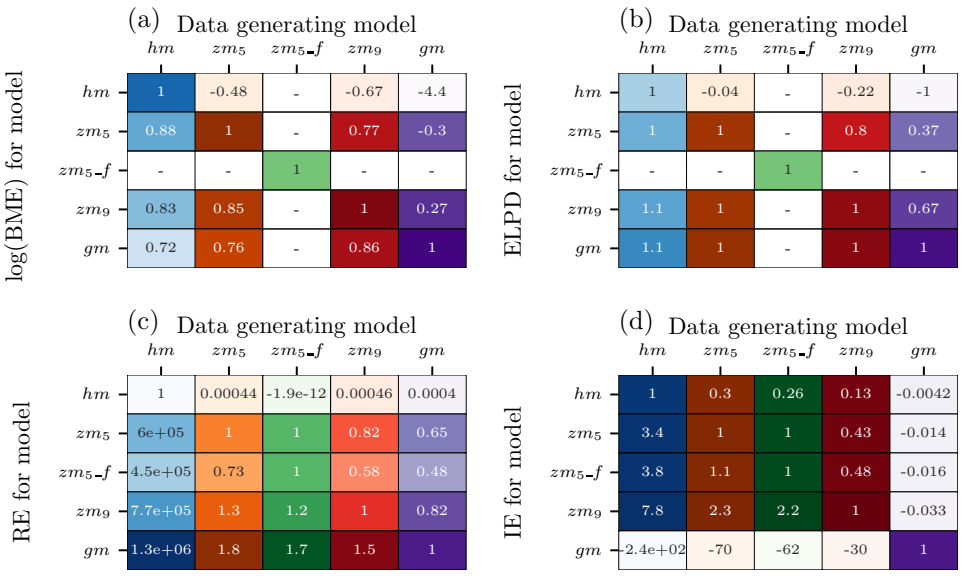

Figure 8. Normalized model confusion matrices for a) ln(BME), b) ELPD, c) RE and d) IE for the Bayesian model similarity analysis. The off-diagonal values are normalized based on the diagonal values. The empty cells in a) and b) correspond to the cases where the flow model is compared against a flow-and-transport model (comparison between models with data sets of different sizes) and thus the MCM entries cannot be used within the similarity analysis.

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. Here, model flexibility, which is visible in the ln(BME) results, has a smaller effect. For example, the differences between the 5 and 9-zoned models seems to has 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

To compare all five models, including the flow model, one can refer to the RE-based MCM in Figure 8c. These results 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. Observe that the normalized RE values are far from one along the first row, independent of M_j. They indicate the inability of the homogeneous model to reproduce all other models, in either the prior, posterior or both. This alludes to large differences in the way the homogeneous model learns from the others. Therefore, if one only had the RE confusion matrix to compare with, one would reach the same conclusion as before: that the homogeneous model is very different from the other models.

When comparing the two 5-zoned models to each other, we can observe that they do not receive the same score when the transport model generates the data. This can be justified by the transport model being able to learn more from 10 reproducible observations than from only five. 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 have similar missing information, which can be supplied by the five additional observations.

On the other hand, the geostatistical model presents the off-diagonal normalized 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.

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 9d in Appendix B shows that there is little to no variability in the score for each model, independent of which model is generating the data. As with BMS, IE induces bias when comparing models with different parameter dimensions. Therefore, when comparing models with dif-

ferent number of uncertain parameters, the IE MCM compares models based solely on the uncertainty induced by the prior parameter distribution.

The two 5-zoned models in Figure 8d present little differences between their IE scores. Both models count with the same prior uncertainty (CE), and therefore the results provide information on the 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 their posterior parameter uncertainty. The results support the previous statement that IE should be compared alongside the other scores, especially RE in case of comparing models with different data sets, and when the parameter dimension between the models is equivalent.

To sum up, through the application to our groundwater models, we addressed issues related to the BME-based approaches for BMS and BMJ. We suggest to complement these frameworks with additional information-theoretic scores to provide a richer picture within the Bayesian framework. We show how these scores can be interpreted in addition to BME. Also, they come at little to no additional computing cost, given that the most computationally demanding step involves the multiple (N_{MC}) model evaluations. We also show the limitations of likelihood-based scores through the inclusion of a competing model with a different data set, as well as how RE and IE can help in overcoming this, at the cost of limiting the comparison to an updating (combined prior and posterior) and posterior uncertainty perspective.

6 Summary and Conclusions

In this study we present how information-theoretic scores, namely expected log predictive density (ELPD), relative entropy (RE) and information entropy (IE), can be used to complement the Bayesian model evidence (BME) for model selection and model justifiably analysis. Employing the connection between Bayesian inference and information theory, we demonstrate how ELPD, RE and IE allow to gain additional insight with regards to 1) posterior model fit (ELPD), 2) information gain in the Bayesian updating (RE) and 3) remaining posterior parameter uncertainty (IE).

We test the proposed methodology on a controlled setup made up of five 2D-groundwater models. These five models each consider a different spatial hydraulic conductivity distribution, which results in different model flexibility. Additionally, we consider both trans-

port and flow models to test the comparison of models with different data sets. For the BMS analysis, the models were compared against a random realization of one of the competing models, with the goal of knowing beforehand the true parameter set and measurement data. For the model similarity analysis, the models were compared against the results from each other.

Arguing based on the mathematical definitions, we show how both BME and ELPD are not suited to compare models with different data sets, neither for BMS nor model similarity analysis. This is due to the bias injected by the normalization factor in the likelihood function, which in this case was considered as a multivariate Gaussian distribution. On the other hand, RE and IE overcome these limitations. One can see this in their definition as information scores for parameter (not data) distributions. When composing them from other ingredients of Bayesian updating we can see that the normalization factor cancels out. Thus they provide a way to compare and select among models in this situation.

In the case of RE, its use with different data sets comes at the cost of solely ranking and comparing among models based on how useful the data was to them, i.e. 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, when not much Bayesian updating is necessary.

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. This can lead to biased results if used on its own to compare models, given that it can eliminate models with a high RE but very uncertain/uninformative prior, or overestimate the appropriateness of a model due to a very simple prior. Therefore, information entropy is useful to complement RE scores, but not as a measure on its own.

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. This additional information can be used by the modeler

to make a better-informed decision based on different perspectives, considering the model setup and the overall modelling goals.

Appendix A: Proofs of the Effect of Data Sets

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 (19) for RE. Here, we use NF to group the normalization factor in the likelihood function (equation (2)), such that:

$$NF = (2\pi)^{\frac{-N_o}{2}} |\mathbf{R}|^{-1/2}. \tag{28}$$

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

$$(\mathbf{y}_k - \mathbf{y}_o) = \boldsymbol{\delta}. \tag{29}$$

Equation (30) shows the simplification of the ln(BME) term based on equation (4):

$$\begin{aligned} -\ln(BME) &= -\ln\left(\mathbb{E}_{prior}\left[NF \cdot \exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) \\ &= -\ln(\mathbb{E}_{prior}[NF]) - \ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) \\ &= -\ln(NF) - \ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) \end{aligned} \tag{30}$$

and equation (31) shows the simplification of ELPD from equation (16) into its basic components:

$$\begin{aligned} ELPD &= \mathbb{E}_{post}\left[\ln\left(NF \cdot \exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right)\right] \\ &= \mathbb{E}_{post}[\ln(NF)] + \mathbb{E}_{post}\left[\ln\left(\exp\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right)\right] \\ &= \ln(NF) + \mathbb{E}_{post}\left(\left(-0.5 \cdot \left[\boldsymbol{\delta}^T \cdot \mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right) \end{aligned} \tag{31}$$

As can be seen in equations (30) and (31), 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 (30) and (31), one can rewrite the equation for relative entropy, based on equation (19), as follows:

$$
\begin{aligned}
RE &= \left(-\ln(NF) - \ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5\cdot\left[\boldsymbol{\delta}^T\cdot\mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right)\right) \\
&\quad + \left(\ln(NF) + \mathbb{E}_{post}\left[\left(-0.5\cdot\left[\boldsymbol{\delta}^T\cdot\mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) \\
&= -\mathbf{ln(NF)} - \ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5\cdot\left[\boldsymbol{\delta}^T\cdot\mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) + \mathbf{ln(NF)} \\
&\quad + \mathbb{E}_{post}\left[\left(-0.5\cdot\left[\boldsymbol{\delta}^T\cdot\mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right] \\
&= -\ln\left(\mathbb{E}_{prior}\left[\exp\left(-0.5\cdot\left[\boldsymbol{\delta}^T\cdot\mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]\right) + \mathbb{E}_{post}\left[\left(-0.5\cdot\left[\boldsymbol{\delta}^T\cdot\mathbf{R}^{-1}\boldsymbol{\delta}\right]\right)\right]. \quad (32)
\end{aligned}
$$

In equation (32) the dependence on the normalization factor NF from both BME and ELPD is canceled out, since it is 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 9 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 Figure 9 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.

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 with 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 - EXC2075 – 390740016 under Germany's Excellence Strategy. We acknowledge the support by the Stuttgart Cen-

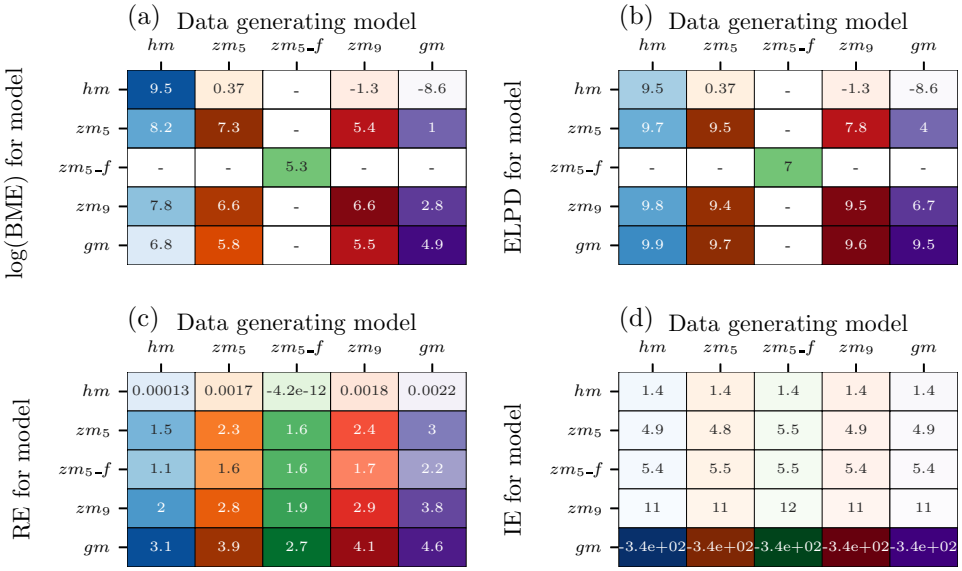

Figure 9. Model confusion matrices for a) ln(BME), b) ELPD, c) RE and d) IE for the Bayesian model similarity analysis. The empty cells in a) and b) correspond to the cases where the flow model is compared against a flow-and-transport model (comparison between models with data sets of different sizes) and thus the MCM entries cannot be used within the similarity analysis

.

945 ter for Simulation Science (SimTech). We would also like to thank the Bundesgesellschaft
946 für Endlagerung (BGE, Federal Company for Radioactive Waste Disposal) for their sup-
947 port.

948 References

949 Akaike, H. (1974). A new look at the statistical model identification. *IEEE transac-*
950 *tions on automatic control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705
951 Aldrich, J. (1997). RA fisher and the making of maximum likelihood 1912-1922. *Sta-*
952 *tistical science*, *12*(3), 162–176. doi: https://doi.org/10.1214/ss/1030037906
953 Altman, N., & Krzywinski, M. (2018). The curse (s) of dimensionality. *Nat Methods*,
954 *15*(6), 399–400. doi: https://doi.org/10.1038/s41592-018-0019-x
955 Bredehoeft, J. (2005). The conceptualization model problem—surprise. *Hydrogeology*

journal, *13*(1), 37–46. doi: 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. doi: 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.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., & Stine, R. A. (2001). The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, *38*, 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. doi: https://doi.org/10.3390/e20020051

Commenges, D. (2015). *Information theory and statistics: an overview.* doi: 10 .48550/ARXIV.1511.00860

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 45–97. doi: 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. doi: 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. doi: 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 (Methodological)*, *56*(3), 501–514. doi: 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. doi: 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. doi: https://doi.org/10.1007/s11222-013-9416-2

Good, I. (1956). Some terminology and notation in information theory. *Proceedings of the IEE-Part C: Monographs*, *103*(3), 200–204. doi: 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). doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1109/TPAMI.1982.4767213

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 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. doi: https://www.jstor.org/stable/2236703

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. doi: 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. doi: https://doi.org/10.1016/j.econlet.2011.12.034

Marsh, C. (2013). Introduction to continuous entropy. *Department of Computer Science, Princeton University*.

Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. *Water Resources Research*, *41*(10). doi: 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. doi: https://doi.org/10.1016/j.advwatres.2018.05.007

Moore, C., & Doherty, J. (2005). Role of the calibration process in reducing model predictive error. *Water Resources Research*, *41*(5). doi: https://doi.org/10.1029/2004WR003501

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, September). *Input data for Bayesian and information theoretic model selection and similarity analysis* [Dataset]. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7086127` doi: 10.5281/zenodo.7086127

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. doi: 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. doi: 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. doi: https://doi.org/10.1007/s00477-003-0151-7

Neuman, S., Wierenga, P. J., & Nicholson, T. (2003). *A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites.* Division of Systems Analysis and Regulatory Effectiveness, Office of Nuclear ....

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(1), 3–26. doi: 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.* 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). doi: 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. doi: 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. doi: 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. doi: https://doi.org/10.3390/e21111081

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

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

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. doi: https://doi.org/10.1016/j.advwatres.2005.11.013

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. doi: https://doi.org/10.1016/j.envsoft.2007.02.004

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. doi: 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). doi: 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. doi: https://doi.org/10.1029/2008WR006908

Santamaría-Bonfil, G., Fernández, N., & Gershenson, C. (2016). Measuring the complexity of continuous distributions. *Entropy*, *18*(3), 72. doi: 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. doi: 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*, 1899-1917. doi: https://doi.org/10.1007/s10596-021-10076-9

Schöniger, A. (2010). *Parameter estimation by ensemble Kalman filters with trans-*

formed data.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464. doi: https://www.jstor.org/stable/2958889

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423. doi: 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. doi: 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.

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 (Methodological)*, *36*(2), 111–133. doi: https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1). doi: 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. doi: 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. doi: 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. doi: 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. doi: 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). doi: https://doi.org/10.1029/2008WR006908

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