The Four Ways to Consider Measurement Noise in Bayesian Model Selection—And Which One to Choose

Sebastian Reuschen1, Wolfgang Nowak1, and Anneli Guthke1,2

1Department of Stochastic Simulation and Safety Research for Hydrosystems, University of Stuttgart, Stuttgart, Germany. 2Now at Stuttgart Center for Simulation Science, University of Stuttgart, Stuttgart, Germany

Abstract Bayesian model selection (BMS) is a statistically rigorous approach to assess the plausibility of competing models. It naturally accounts for uncertainties in models and data. In this study, we discuss the role of measurement noise in BMS deeper than in past literature. We distinguish between four cases, accounting for noise in models and/or data: (1) no-no, (2) no-yes, (3) yes-no, and (4) yes-yes. These cases differ mathematically and philosophically. Only two out of these four cases are logically consistent, and they represent two potentially conflicting research questions: "Which model is best in modeling the pure physics?" (Case 1) and "which model is best in predicting the data-generating process (i.e., physics plus noise)?" (Case 4). If we are interested in the "pure physics question," we face two practical challenges: First, we would need noise-free data, which is impossible to obtain; and second, the numerical approximation of Bayesian model evidence can be hard when neglecting noise. We discuss how to address both challenges and reveal that a fallback to the easier "data-generation question" as a proxy for the "physics question" is not appropriate. We demonstrate on synthetic scenarios and a real-world hydrogeological case study that the choice of the case has a significant impact on the outcome of posterior model weights, and hence on results of the model ranking, model selection, model averaging, model confusion analysis, and uncertainty quantification. Reality might force us to use a different case than philosophy would suggest, and we provide guidance on how to interpret model probabilities under such conditions.

1. Introduction

Models are used to predict and/or investigate and explain phenomena in nature. Often, many hypotheses exist for these two tasks. Naturally, the question arises, which of the competing modeling approaches predicts or explains nature best. Bayesian model selection (BMS, e.g., Wasserman, 2000) is a statistical method that uses observed data to select between competing models. BMS is settled in a rigorous probabilistic framework and follows the scheme of Bayesian updating: A prior belief about the plausibility of each candidate model is updated to a posterior model weight in the light of measured data (i.e., the probability of the model to have generated the data, given the model set). Posterior model weights are then used as a basis for Bayesian model ranking, selection, or averaging (BMA, Hoeting et al., 1999).

To help with the interpretation of posterior model weights, the so-called model confusion matrix (MCM) has been introduced by Schöniger, Illman, et al. (2015). It reveals whether a lack of confidence in model choice is due to similarity between the candidate models or due to weakly informative data. The MCM is a purely synthetic analysis that can be used as a scale of reference for model weights obtained from real data. Schäfer Rodrigues Silva et al. (2020) have recently extended the MCM analysis to identify the best surrogate model from a set of candidates to replace an expensive full-complexity model in stochastic analysis.

Technically, the Bayesian updating procedure requires calculating the so-called Bayesian model evidence (BME). BME is the likelihood of a model to have generated the data, integrated over its whole parameter space and all involved probability distributions. While the likelihood accounts for uncertainty in measured data, the integration considers parameter uncertainty, and potentially also uncertainty in model drivers or boundary conditions. In some cases, the integration even accounts for statistical representations of model errors (Leube et al., 2012; Nowak et al., 2012), which is perceived by many studies to be part of the likelihood. Due to its statistical rigor and its elegance in accounting for uncertainty, BMS has become popular in water resources research. It has been applied in various different contexts, such as evaluation of hydrologi-
We argue that equipping models with measurement noise is of course logically consistent, but answers a research question that is not at the core scientific interest: “which model is best in predicting the data-generating process (i.e., physics plus noise)?” The more intriguing question would be “which model is best in modeling the pure underlying physics?” (in the spirit of Bayesian hypothesis testing), and this question could only be addressed by comparing noise-free model (\( M_k \)) predictions with noise-free data (\( d_0 \)). This is hindered by two practical challenges: First, we do not have noise-free observation data (we call this the “data-availability barrier”); and second, numerical approximation of BME can become an intractable problem with the likelihood function collapsing to a Dirac delta function (we call this the “numerical-approximation barrier”). We comment on these challenges in Sections 3.4 and 3.5.

We put these issues into a consistent frame by enumerating all four combinatorially possible ways of how to treat noise in BMS: First, we can assume data to be noise-free, and predict with noise-free models (Case 1). Second, we may believe to have noisy data but choose not to simulate it with our models (Case 2). Third, we can add noise to our predictions but compare them to (allegedly) noise-free data (Case 3). And fourth, we can consider noise in both models and data (Case 4). This is illustrated by the “decision matrix” in Figure 1.

While the ability to account for uncertainties is the primary reason for the popularity of the Bayesian framework in hydrological multi-modeling analyses, the theoretical basis of how to treat measurement noise in BMS has rarely been in the focus. Lu et al. (2013) have investigated the effect of error covariance structure in the likelihood function on posterior weights for transport models. Schöninger, Wöhling, and Nowak (2015) have demonstrated that posterior weights of soil-plant-atmosphere models can vary significantly under random outcomes of measurement noise in the observed data and that this variability can even change model ranking and corresponding modeling conclusions. What is missing so far is a theoretical dissection of how and for which reasons measurement noise should be treated in BMS. We wish to make modelers aware of the impact this treatment has on the final outcome of model ranking results and dependent conclusions.

For our discussion, we assume a deterministic “physics model” \( f_k(u_k, v_k) \) that produces probabilistic output, for example, due to uncertain parameters \( u_k \) and/or inputs \( v_k \). The modeled output is potentially complemented by a statistical model error term \( \varepsilon_{M_k} \). The result is a prediction of the true state (without measurement noise) and we call the corresponding hypothetical observation-noise-free data \( d_0 \). To make predictions of real (measurement noise-corrupted) noisy data \( \hat{d}_i \), a measurement noise term \( \varepsilon_{d_i} \) is added to the physics prediction:

\[
\hat{d}_i = f_k(u_k, v_k) + \varepsilon_{M_k} + \varepsilon_{d_i} + d_0
\]

Equation 1 shows that we can either predict physics only (denoted as model \( M_k \)), or we can predict the data-generating process (physics + noise, denoted as model \( \hat{M}_k \)). We argue that equipping models with noise (\( M_k \)) and comparing them to noisy data (\( \hat{d}_i \)) is of course logically consistent, but answers a research question that is not at the core scientific interest: “which model is best in predicting the data-generating process (physics + noise)?” The more intriguing question would be “which model is best in modeling the pure underlying physics?” (in the spirit of Bayesian hypothesis testing), and this question could only be addressed by comparing noise-free model (\( M_k \)) predictions with noise-free data (\( d_0 \)). This is hindered by two practical challenges: First, we do not have noise-free observation data (we call this the “data-availability barrier”); and second, numerical approximation of BME can become an intractable problem with the likelihood function collapsing to a Dirac delta function (we call this the “numerical-approximation barrier”). We comment on these challenges in Sections 3.4 and 3.5.

We put these issues into a consistent frame by enumerating all four combinatorially possible ways of how to treat noise in BMS: First, we can assume data to be noise-free, and predict with noise-free models (Case 1). Second, we may believe to have noisy data but choose not to simulate it with our models (Case 2). Third, we can add noise to our predictions but compare them to (allegedly) noise-free data (Case 3). And fourth, we can consider noise in both models and data (Case 4). This is illustrated by the “decision matrix” in Figure 1.

We shed light on the four cases from a philosophical, mathematical, and numerical perspective, and discuss ways to overcome or deal with these barriers. We focus on the two logically consistent cases (cf. Figure 1) that represent the two potentially conflicting research questions mentioned above and investigate whether we can exploit the inconsistent cases to approximate the pure-physics case.

We show that the theoretical differences between the four cases translate into differences in model weight outcomes. We test whether these differences matter in real-world applications (i.e., whether there is a risk that the model identified as the “most plausible physics model” is different from the identified “most plausible data-generating model”). To do so, we investigate in three analytical test case scenarios under which conditions the differences in model weights are most pronounced. We also derive a mathematical
formulation of model confusion weights for all four cases and use the MCM results as a tool to interpret the severity of differences between the cases. Finally, we demonstrate the implications of the four cases on a real-world hydrogeological case study that features data from a sandbox aquifer lab experiment by Illman et al. (2010).

We here investigate how choosing between our four cases impacts an example of BMS for a set of groundwater models in a brute-force Monte Carlo setup. Yet, our theoretical discussion is general enough to be also applicable (a) to other multi-model approaches such as BMA, pseudo-BMA, or Bayesian stacking (e.g., Höge et al., 2020), to (b) other numerical implementation schemes (e.g., Liu et al., 2016; Volpi et al., 2017) or mathematical approximations of model weights via information criteria (e.g., Schöniger et al., 2014; Ye et al., 2008), and (c) to arbitrary other fields of applications in water resources research and beyond.

The main contributions of this study are the following:

1. We discuss where and for which reasons measurement noise should be considered in BMS (four cases).
2. We distinguish between two potentially conflicting research questions and identify the cases in which they can be pursued in a logically consistent way.
3. We reveal practical hindrances and analyze which approximations can be taken to still approach the desired research question.
4. We finally provide a recipe for BMS that ensures philosophically consistent and numerically stable results under noise.

We summarize the existing mathematical framework of BMS, including the derivation of the MCM, in Section 2. In Section 3, we introduce the four different ways to handle measurement noise and discuss them from a philosophical, mathematical, and practical perspective. Section 4 demonstrates the differences between these cases on simplified 1D analytical examples. Then, we present the application of the four cases to a real-world hydrogeological case study in Section 5. Finally, we summarize findings and provide a recipe for BMS that ensures philosophically consistent and numerically stable results under noise in Section 6.

2. Bayesian Model Selection Framework

2.1. Bayesian Model Selection

BMS is a statistical framework to choose between competing models (Raftery, 1995). It compares the predictive distributions of the different models and ranks the models in their ability to predict the measured data. Besides the goodness-of-fit, BMS takes into account the complexity of the models. It, therefore, yields a model ranking that reflects a compromise between performance and parsimony. A general introduction to the BMS framework can be found in Höge et al. (2019).
When using the BMS framework, we implicitly assume that we are in an M-closed setting (Bernardo & Smith, 2009). That is, we assume the data was produced by one of the \( N_m \) considered models (and hence, given the parameters \( \int \)). This probability is also called and call it. Based on the data produced by model \( M_k \) in model \( M_k \), which is called Bayesian model denoting probability densities of continuous random variables, and \( \text{uni} \). This probability to be the true one. Represents the prior belief in model \( \frac{1}{2} \). \( k \) (of model \( k \)-th column of the \( k \)-th row and \( k \)) represents denoting discrete probability distributions. The prior model weight \( \text{uni} \). This prior belief is updated by the predicted density of the data, \( \text{uni} \), which is called Bayesian model evidence (BME). It is defined as the expected likelihood of the observed data integrated over the model’s prior parameter space \( \text{uni} \):

\[
P(d_0|M_k) = \int_{\text{uni}} p(d_0|M_k, u_k) p(u_k|M_k) \, du_k
\]

(3)

Here, the prior distribution of parameters \( u_k \) in model \( M_k \) is denoted by \( p(u_k|M_k) \). \( p(d_0|M_k, u_k) \) represents the likelihood of the data \( d_0 \) given the parameters \( u_k \) of model \( M_k \).

In practice, we are missing a reference to compare the obtained Bayesian model weights to, in order to judge their decisiveness. One way to move forward is to use the so-called Bayes factor (Kass & Raftery, 1995) to determine the decisiveness of evidence in favor of one out of two models; however, the Bayes factor does not tell us how decisive this choice could ideally be under the given conditions (model set, uncertainty in models, and available data). This is where the model justifiability analysis proposed by Schöniger, Illman, et al. (2015) comes into play: it determines in a synthetic setting, how decisive model weights could be at most (i.e., if any of the models in the set was actually the true one). The core ingredient of the model justifiability analysis is the model confusion matrix, which will be presented in the next Section.

2.2. Model Confusion Matrix

The model confusion matrix (MCM) was proposed by Schöniger, Illman, et al. (2015) as the basis for a so-called model justifiability analysis. The MCM consists of posterior model weights for the models in the set under the condition that the data set was actually produced, in turn, by each one of the models. Since we assume model predictions to be probabilistic (due to, e.g., parameter and input uncertainty), there is not a single data set representative of each model, but instead, we have to consider the models’ predictive distribution. That means, the given data set \( d_0 \) from Equations 2 and 3 now turns into a random number \( d_\ell \) with distribution \( p(d|M_\ell), \ell = 1 \ldots N_m \).

As we now fulfill the requirement of an M-closed setting, we obtain Bayesian probabilities under ideal conditions. Within this synthetic setting, we can exclude noise, model errors, and other annoyances from the analysis, and put the focus on the theoretical information content of the data set. This information is characterized by the type of data and the data set length; or, more specifically: it is given by the sensitivity of model choice to the experimental setup. The goal of the justifiability analysis is to identify an upper limit to the decisiveness of model weights, which can be used to judge the model ranking result obtained for the real data set \( d_0 \).

The columns of the MCM represent the data-generating model, whereas the rows list the models to be evaluated, that is, that aim to predict the given data. To populate the \( k \)-th row and \( \ell \)-th column of the MCM, we determine the expected weight for model \( M_k \) based on the data produced by model \( M_\ell \) and call it \( P(M_k|M_\ell) \). This expected posterior model weight is given by

\[
P(M_k|M_\ell) = E_{\tilde{D}} \left[ P(M_k|\tilde{d}) \right] = \int_{\tilde{D}} P(M_k|\tilde{d}) p(\tilde{d}|M_\ell) \, d\tilde{d}
\]

\[
= \frac{\int_{\tilde{D}} p(\tilde{d}|M_k)p(M_k)}{\sum_{i=1}^{N_m} \int_{\tilde{D}} p(\tilde{d}|M_i)p(M_i)} \, d\tilde{d}
\]

(4)
where \( p(\mathbf{d}|M_r) \) denotes the distribution of synthetic data generated by model \( M_r \). In the following, we assume that prior model weights are uniformly distributed (all \( P(M_i) \) are identical), which simplifies Equation 4 to:

\[
P(M_k|M_r) = \mathbb{E}_D \left[ \frac{p(\mathbf{d}|M_k)}{\sum_{i=1}^{N_m} p(\mathbf{d}|M_i)} \right] = \int_D \frac{p(\mathbf{d}|M_k)}{\sum_{i=1}^{N_m} p(\mathbf{d}|M_i)} p(\mathbf{d}|M_r) \, d\mathbf{d}.
\] (5)

The MCM provides two major insights. First, the main diagonal reveals the maximum self-identification weights given the experimental setup. These weights can be used as a reference for BMS results with real data: the diagonal values \( P(\mathbf{M}_k, \mathbf{M}_r) \) tell us how large the model weight for model \( \mathbf{M}_1 \) would be if it perfectly represented nature. Hence, if our BMS analysis resulted in a posterior weight of 60% for model 1 and \( P(\mathbf{M}_1, \mathbf{M}_1) = 0.6 \), we would know that we could not get any higher weight (on average) under the given experimental setup. We would conclude that (a) based on the observed data model 1 most probably has produced the data and (b) our measurement data is not well suited to distinguish between the competing models with high confidence. One way to solve the latter issue is to take more or more informative (i.e., more precise or more sensitive) measurements, which will increase the values on the main diagonal of the MCM and decrease the values on the off-diagonals.

Second, the off-diagonal entries reveal the similarity between models. The value \( P(\mathbf{M}_k, \mathbf{M}_r) \) indicates how similar model \( \mathbf{M}_k \) and model \( \mathbf{M}_r \) are under the available data. Here, large values represent a high predictive similarity of the models. If the value is as large as the diagonal entry of, for example, \( P(\mathbf{M}_k, \mathbf{M}_k) \), it follows that models \( \mathbf{M}_k \) and \( \mathbf{M}_r \) cannot be distinguished based on this specific experimental setup. Again, taking more measurements might solve this problem. A value of zero on the off-diagonals represents perfectly distinguishable models with no similarity at all. The similarity of models is relevant, for example, identifying the best surrogate model for a (too) expensive high-fidelity model. To this end, the column with the data of the high-fidelity model is used to identify the most similar surrogate model, that is, the model showing the highest confusion with the high-fidelity model (Schäfer Rodrigues Silva et al., 2020).

3. The Four Ways to Consider/Ignore Noise in Bayesian Model Selection

As laid out in Sections 1 and 2, the Bayesian framework for model evaluation is designed to account for uncertainty in models and data. Here, we focus on measurement uncertainty. Typically, measurement noise is accounted for in the likelihood function. We will explain that this approach is equivalent to adding noise to model predictions and that this is only one out of four possible perspectives on the noise handling theme.

Here and in the following, we distinguish between hypothetical noise-free data \( \mathbf{d} \) and real noisy data \( \mathbf{d}_0 \). Noise-free data refers to data that represents the true system state. Noisy data, in contrast, refers to observations of the system state that suffer from measurement noise. Of course, “measurement noise” is a conceptual simplification with a frequentist interpretation: to us, repeated measurements of the same system state lead to a set of outcomes that center on the true system state (assuming an unbiased observation procedure) and that shows a certain level of variability (= noise). Similarly, our models \( \mathbf{M}_k \) target the true system state, and we can add a noise description to obtain models \( \tilde{\mathbf{M}}_k \) that aim to reproduce the expected value and the variability in the data.

We shed light on the four cases defined by Figure 1 from a scientific/philosophical perspective in Section 3.1. Then, we provide a rigorous mathematical derivation of posterior model weights and model confusion weights in the four cases in Sections 3.2 and 3.3, respectively. Section 3.4 discusses challenges in numerical implementation, and Section 3.5 discusses constraints posed by limited data availability. Finally, in Section 3.6, we combine philosophical, mathematical, and pragmatic considerations to briefly summarize implications for model selection in the presence of measurement noise.

3.1. Philosophical Perspective

The different cases are shown in Figure 1 model four different scenarios as schematically illustrated in Figure 2.
3.1.1. Case 1

Figure 2 illustrates that, in Case 1, the model simulates the physical processes that cause the (only theoretically observable) true system state; these processes are summarized as “physics.” If this true system state was observable, we could compare our predictions with noise-free data to judge the model’s quality. Such comparison would be fair and consistent, and very insightful if we wanted to learn about the true processes acting in the (hydrological) system. A typical application of Case 1 is to identify the most plausible physical model, that is, to increase system understanding by identifying dominant processes.

3.1.2. Case 2

In Case 2, the model shall represent the true system state by simulating physics only; but the model output is compared to real noisy data. This case is logically inconsistent because model output and data do not match: the “noise-generating” observation process is missing in the model formulation. We label this case as inconsistent because noise-free predictions are compared to noisy data. Hence, there is no philosophical reason to use this case.

3.1.3. Case 3

Case 3 models what we call the “data-generating process” in Figure 2 and in the following. This includes modeling both the physics and the (simplified) observation procedure that introduces measurement noise. Using noise-free data for evaluating such a model produces a conceptual error, and hence we label this case as inconsistent as well. Again, there is no philosophical reason to use this case.

3.1.4. Case 4

Case 4 models the data-generating process and uses noisy data for model-data comparison. This case is hence logically consistent. Similarly, Nearing et al. (2016) argue that a set of modeling hypotheses can only be tested as a whole, i.e., by modeling the full measurement (error) process. In our words, they argue that only the “data-generating process” question can be answered. Hence, in Case 4, we cannot assess the plausibility of the physics model alone. A typical application of Case 4 could be, for example, predicting whether a certain legal threshold value will be exceeded - here it is only relevant to predict the combined signal of physics and observation process; the true system state cannot be uncovered and hence cannot be the subject of regulatory guidelines.

3.2. Mathematical Definition of Posterior Model Weights

In the specific context of Bayesian model selection, we wish to formalize the impact of the four introduced cases on the calculation of posterior model weights, assuming noise-free or noisy model predictions and given noise-free or noisy data.

In the noise-free data case, we simply evaluate posterior model weights once on the given data $d_0$. In contrast, evaluating posterior model weights given real noisy data is more challenging. What we do in practice is to determine model weights once, given the observed noisy data $\tilde{d}_0$. This data contain a specific but unknown realization of noise $e_0$ (note that we drop the subscript $d$ when describing $e$ in the remainder of this work), with $\tilde{d}_0 = d_0 + e_0$. Due to the conceptualization of noise $e_0$ as a realization of a random variable, also noisy data $\tilde{d}$ becomes a random variable $\tilde{d} = d_0 + e$. Note that we assign fixed given values a subscript ($d_0, e_0$), while random numbers are written without subscript ($d, e$). The random variable $\tilde{d}$ reveals its randomness, e.g., under repetition of an observation or experiment. The distribution of $\tilde{d}$ is given by

$$p(\tilde{d}) = \delta_{d_0} * p_e$$

where $*$ denotes a convolution, $\delta_{d_0}$ is the probability density function of $d$ (a Dirac delta function centered at $d_0$), and $p_e$ is the measurement noise distribution.
To investigate the general differences between the four cases, we integrate over all possible outcomes of random noisy data \( \hat{d} \). This means that we compare expected values of posterior model weights given noisy data in Cases 2 and 4 with fixed posterior model weights given noise-free data in Cases 1 and 3. For consistency, we formulate all results as expected values over \( p_E \): in Cases 1 and 3, the expected value is over a Dirac delta noise function centered at 0 (\( p(d) = \delta_{d_0} \ast p_E = \delta_{d_0} \ast \delta = \delta_{d_0} \)), which collapses to the result given \( d_0 \); in Cases 2 and 4, the expectation is over the random part of \( \hat{d} \), that is, over \( \varepsilon \).

### 3.2.1. Case 1

Case 1 compares predictions of the true system state by model \( M_k \) with noise-free data \( d_0 \). Not modeling the noise in predictions and having noise-free data leads to posterior model weights

\[
P(M_k | d_0) = \sum_{i=1}^{N_m} p(d_0 | M_i),
\]

still assuming uniform prior model weights to simplify the equations.

As stated above, in the noise-free data case, the expected posterior model weight over \( p(d) \) is equal to the posterior model weight given \( d_0 \)

\[
\mathbb{E}_E [P(M_k | d)] = P(M_k | d_0).
\]

### 3.2.2. Case 2

In Case 2, we compare predictions of the true system state by model \( M_k \) with noisy data \( \hat{d}_0 \) to obtain the posterior model weight \( P(M_k | \hat{d}_0) \):

\[
P(M_k | \hat{d}_0) = \sum_{i=1}^{N_m} p(\hat{d}_0 | M_i),
\]

Since the random outcomes of \( \hat{d} \) are defined by the deterministic (but unknown) true \( d_0 \) and random outcomes of noise \( \varepsilon \), we can use the convolution formulation (cf. Equation 6) to express the expected posterior model weight as a function of \( p_E \):

\[
\mathbb{E}_E \left[ P(M_k | \hat{d}) \right] = \mathbb{E}_E \left[ P(M_k | d) \right] = \int_{-\infty}^{\infty} \frac{p(d | M_k)}{\sum_{i=1}^{N_m} p(d | M_i)} \cdot p(d) \, d\hat{d} = \int_{-\infty}^{\infty} \frac{p(d_0 + \varepsilon | M_k)}{\sum_{i=1}^{N_m} p(d_0 + \varepsilon | M_i)} \cdot p_E(\varepsilon),
\]

### 3.2.3. Case 3

In Case 3, we predict the noisy system state with models \( \hat{M}_k \), but use noise-free data \( d_0 \) for model evaluation. This leads to

\[
P(\hat{M}_k | d_0) = \sum_{i=1}^{N_m} \frac{p(d_0 | \hat{M}_i)}{p(d_0 | M_i)} = \frac{p(d_0 | M_k) \ast p_E}{\sum_{i=1}^{N_m} (p(d_0 | M_i) \ast p_E)},
\]

using the fact that the predictive PDF produced by \( \hat{M}_k \) can be obtained by widening the noise-free predictive PDF by \( M_k \) via convolution with \( p_E \).

Like in Case 1, we formally consider the noise distribution defining the outcome of \( d \) as a Dirac delta function at \( \varepsilon_0 = 0 \), which leads to

\[
\mathbb{E}_E \left[ P(\hat{M}_k | d) \right] = P(\hat{M}_k | d_0).
\]

One might think that considering noise either in the data (Case 2) or in the model (Case 3) would lead to the same results. This is true for the calculation of the expected BME (\( \mathbb{E}_E \left[ p(\hat{d} | M_k) \right] = p(\hat{d} | M_k) \ast p_E = \mathbb{E}_E \left[ p(\hat{d} | \hat{M}_k) \right] \)). However, when targeting expected posterior model weights, the expectation is not over BME. Rather, model weights are obtained by normalization with the sum of all BMEs in the denominator before the
expectation is taken. This is why expected posterior model weights differ between Cases 2 and 3 ($E\varepsilon \left[ P(M_k|\hat{d}) \right] \neq E\varepsilon \left[ P(\hat{M}_k|d) \right]$). In Case 2 (Equation 10), the normalization happens per noise realization, while in Case 3 (Equation 11) each model’s predictive PDF is widened by convolution with $p(\epsilon)$ before the normalization.

### 3.2.4. Case 4

In Case 4, we simulate the noisy system state by model $\hat{M}_k$ and compare these predictions to given noisy data $\hat{d}_0$. This leads to posterior model weights:

$$P(M_k|\hat{d}_0) = \frac{p(d_0|M_k) \ast p(\epsilon)}{\sum_{i=1}^{N_m} p(d_0|M_i) \ast p(\epsilon)}. \quad (13)$$

Averaging this over all possible realization of noise leads to:

$$E\varepsilon \left[ P(M_k|\hat{d}) \right] = \frac{p(d_0|M_k) \ast p(\epsilon)}{\sum_{i=1}^{N_m} (p(d_0|M_i) \ast p(\epsilon))} \ast p(\epsilon). \quad (14)$$

### 3.3. Mathematical Definition of Model Confusion Weights

Next, we derive the model confusion weights for the four cases. The only difference to posterior model weights as defined in Section 3.2 is that the data $d_0$ is not given in the form of observations anymore. Instead, we assume that model $M_r$ produces synthetic data $\tilde{d}$ (noise-free) or $\tilde{d} + \epsilon$ (noise-perturbed). Hence, instead of a Dirac delta for $p(d)$, we deal with a distribution of possible synthetic data $p(d|M_r)$ produced by model $M_r$. When accounting for noise, model $\hat{M}_r$ produces noisy synthetic data $p(d|\hat{M}_r) = p(d|M_r) \ast p(\epsilon)$.

#### 3.3.1. Case 1

In Case 1, we use models $M_k$ to predict the noise-free system state and compare against noise-free synthetic data by model $M_r$. Hence, the corresponding expected posterior model weight is given by:

$$P(M_k|M_r) = \int_D \frac{p(\tilde{d}|M_k)}{\sum_{i=1}^{N_m} p(\tilde{d}|M_i)} \, p(\tilde{d}|M_r) \, d\tilde{d}. \quad (15)$$

This equation is identical to the general MCM formulation in Equation 5, with the only difference that we now explicitly specify $\tilde{d}$ to be noise-free (in our general derivation in Section 2.2, data can be noisy or not).

#### 3.3.2. Case 2

Now, while still keeping $M_k$ noise-free, we sample from a model $\hat{M}_r$ that produces noisy data $\hat{d} + \epsilon$ and obtain the posterior model confusion weight:

$$P(M_k|\hat{M}_r) = \int_D \frac{p(\hat{d}|M_k)}{\sum_{i=1}^{N_m} p(\hat{d}|M_i)} \cdot p(\hat{d}|M_r) \ast p(\epsilon) \, d\hat{d}. \quad (16)$$

#### 3.3.3. Case 3

When we evaluate model confusion weights for model $\hat{M}_k$ that considers noise, but do not add noise to the synthetic data produced by model $M_r$, we obtain the following expression:

$$P(\hat{M}_k|M_r) = \int_D \frac{p(\hat{d}|M_k) \ast p(\epsilon)}{\sum_{i=1}^{N_m} p(\hat{d}|M_i) \ast p(\epsilon)} \, d\hat{d}. \quad (17)$$

As for posterior model weights, model confusion weights differ between Cases 2 and 3 due to the normalization in the denominator.

#### 3.3.4. Case 4

In Case 4, we simulate the noisy system state and compare it with noise-perturbed synthetic data. Hence, model confusion weights are given by
Figure 3. Barriers that hinder transitions between the cases: impervious data-availability barrier separating Cases 2 and 4 from 1 and 3 vertically, and semi-pervious numerical-approximation barrier separating Cases 1 and 2 from 3 and 4 horizontally. Red data column is impossible to populate in real data scenarios. Green model row is numerically straightforward to obtain with satisfying accuracy; the red model row relies on rough numerical approximations.

3.4. Numerical Implementation

3.4.1. Bayesian Model Evidence

The numerical challenge lies in approximating the BME value per model, as a basis for posterior model weights and model confusion weights. If a model’s predictive PDF is known analytically, we can determine BME easily as the PDF value at the given data value(s). For most real-world applications, however, BME cannot be evaluated analytically and the integral in Equation 3 needs to be approximated numerically or mathematically instead.

In our study, we will use brute-force Monte Carlo sampling (Gelman et al., 1995, chapter 10) of the prior distributions. As argued by Schöniger et al. (2014), Monte Carlo is superior to other numerical schemes in that it is an unbiased scheme that is known to converge to the correct limit, and its convergence can be easily monitored. In our chosen test cases, the computational burden of Monte Carlo is bearable; for computationally heavier practical applications, alternative numerical methods could be used to improve on computational efficiency, such as nested sampling (Elshiekh et al., 2014; Skilling, 2006), thermodynamic integration (Lartillot & Philippe, 2006; Liu et al., 2016), stepping stone sampling (Elshall & Ye, 2019; Xie et al., 2011), or Gaussian mixture importance sampling (Volpi et al., 2017), to name a few examples. However, these methods are less straightforward to implement and bear the risk of introducing biases into the BME estimation.

The Monte Carlo approximation of BME (exemplarily shown here for Case 3) is based on \( N_{MC} \) random samples (realizations with subscript \( r \)) of the model’s prior parameter space \( U_k^2 \):

\[
p(d_k|M_k) \approx \sum_{i=1}^{N_{MC}} p(d_k|\hat{M}_k, u_{k,r}) = \sum_{i=1}^{N_{MC}} (p(d_0|M_k, u_{i,r}) \ast p_e).
\]

Assuming uncorrelated, unbiased measurement errors (i.e., \( p_e \) being a normal distribution with mean zero and standard deviation \( \sigma \)), the likelihood function \( p(d_0|\hat{M}_k, u_{i,r}) \) can be expressed as:

\[
p(d_0|\hat{M}_k, u_{i,r}) = \frac{1}{\sqrt{(2\pi)^{N_o}\det(R)}} \exp \left( -\frac{1}{2} (d_0 - y_{i,r})^T R^{-1} (d_0 - y_{i,r}) \right),
\]

where \( y_{i,r} \) is the prediction of model \( M_k \) with parameters \( u_{i,r} \). The error covariance matrix \( R \) is of size \( N_o \times N_o \), and features main diagonal entries of \( \sigma^2 \). Other choices of \( p_e \) are possible and would affect the results of individual BME values; investigating whether and how other likelihood functions aggravate the differences between the four cases is beyond the scope of this study.

Approximation of BME with Monte Carlo integration is hence straightforward if we understand the noise description as part of the model (Cases 3 and 4). In contrast, Cases 1 and 2 do not consider measurement noise in predictions. This could be interpreted as a standard deviation \( \sigma \) of zero in Equation 20. Mathematically, this leads to a Dirac delta likelihood function centered about the model prediction \( y_{i,r} \), and practically all parameter realizations would receive a likelihood of zero. Hence, it is practically impossible to approximate BME with a likelihood-based numerical scheme in Cases 1 and 2. We call this the “numerical
approximation barrier” and visualize it in Figure 3. Figure 3 illustrates the practical feasibility of the four cases. We distinguish between challenges regarding numerical approximation (discussed in this section) and regarding data availability (discussed in Section 3.5). The numerical approximation barrier horizontally separates Cases 1 and 2 from Cases 3 and 4. This barrier can be understood as semi-pervious - we can push the barrier toward Cases 1 and 2 by applying a numerical trick: we increase the theoretical standard deviation from $\sigma = 0$, representing zero noise, to a very small value $> 0$. This allows us to obtain a (biased) value of BME by using Monte Carlo integration according to Equation 19. An adaptive algorithm that automatically determines an optimal $\sigma$ for this approximation would be handy and is left for future studies.

### 3.4.2. Posterior Model Weights

Once BME has been determined for all models in the set, the calculation of posterior model weights according to Equation 2 is straightforward. In Cases 2 and 4, we are interested in expected model weights over possible outcomes of $\hat{d}$, that is, over the noise distribution $p_\epsilon(\epsilon)$. Assuming that we hypothetically know the true data value $d_0$, we can approximate the expected value using brute-force Monte Carlo integration over random realizations of noise in three steps:

1. Sample from the measurement noise distribution $p_\epsilon$. We draw $N_d$ samples (realizations) and denote the $j$th sample as $\epsilon_j$.
2. Determine the posterior model weight for each of these data sets: $P(M_k|d_0 + \epsilon_j) = \frac{p(d_0 + \epsilon_j|M_k)}{\sum_{i=1}^{N_m} p(d_0 + \epsilon_i|M_i)}$.
3. Report the average value of the posterior model weights over all data samples $\epsilon_j$: $\mathbb{E}_\epsilon P(M_k|\hat{d}) \approx \frac{1}{N_d} \sum_{j=1}^{N_d} P(M_k|d_0 + \epsilon_j)$.

This example shows the approximation for Case 2 where we use a noise-free model $M_k$. To calculate posterior model weights in Case 4, simply replace $M_k$ by $\hat{M}_k$ in all three steps.

In real-world applications, we are typically given a single realization $d_0$, and $\epsilon_j$ is unknown. Hence, we cannot average over repeated outcomes of $\epsilon$, and rely on the model weight given $d_0$ instead.

### 3.4.3. Model Confusion Weights

For the calculation of MCM entries, we follow the approach proposed by Schöninger, Illman, et al. (2015) and approximate Equation 15 in the same three steps:

1. Sample from each predictive distribution $p_\epsilon(d|M_l)$, $l = 1 \ldots N_m$. We draw $N_d$ samples (realizations) per model and denote the $j$th sample as $d_{l,j}$.
2. Determine the posterior model weight for each of these data sets: $P(M_k|d_{l,j}) = \frac{p(d_{l,j}|M_k)}{\sum_{i=1}^{N_m} p(d_{l,i}|M_i)}$.
3. Report the average value of the posterior model weights over all data samples $d_{l,j}$: $P(M_k|M_l) = \mathbb{E}_{d_{l,j}} [P(M_k|d_{l,j})] \approx \frac{1}{N_d} \sum_{j=1}^{N_d} P(M_k|d_{l,j})$.

This example shows the approximation of Case 1. For Case 3, simply replace $M_k$ by $\hat{M}_k$; for Cases 2 and 4, replace $M_l$ by $\hat{M}_l$, that is, additionally sample the noise distribution as described for posterior model weights above.

### 3.5. Constraints Due to Data Availability

For model selection given a real experimental or field data set, all we have is noisy data, that is, we are bound to the right column of the decision matrix in Figure 1: Cases 2 and 4. We call this constraint the “data-availability barrier” and illustrate its location in Figure 3. This barrier can be understood as impervious—there is no way to remove the noise from the observed data in order to transition to the left column of the decision matrix because we do not know the actual outcome of random measurement noise that is added to the true system state value (and hence, we are unable to identify the exact true system state).

### 3.6. Brief Synthesis of Philosophical, Mathematical, and Pragmatic Aspects

We claim that modelers always want to be in Case 1 or Case 4 because these cases are consistent and do not introduce conceptual errors. We carefully choose the word “want” instead of “should” here, because we might not always be able to choose these cases.
Philosophically, the choice is clear: either one is interested in pure physics, then one chooses Case 1, or one is interested in simulating the data-generating process, then one move to Case 4. If one is interested in identifying the model that has most likely produced the noisy data (that mimics the data-generating process best), one simply has to pay attention to equip the model predictions with noise when calculating model confusion weights. Indeed, this is also computationally the most straightforward way to go.

Case 1 is only meaningful if noise-free data are available, for example, when aiming to identify the most suitable surrogate model for a given reference model. As soon as real noisy data are involved as the basis for determining model weights, we are faced with the data-availability barrier that forces us to stay with Cases 2 and 4.

In hydro(geo)logical applications, we are additionally faced with the numerical-approximation barrier, because model output PDFs are typically not given analytically. This leaves us with a straightforward implementation of Cases 3 and 4; and the barrier can be pushed to approximate Cases 1 and 2 by introducing a “small enough” noise level into the likelihood function.

From the mathematical formulation of the four cases, we notice that the MCMs of Case 1 and Case 4 are symmetric. It is interesting to note that both consistent cases are symmetric, while the inconsistent cases are not. Consequently, we always want to obtain a symmetric MCM; any asymmetry of the MCM indicates either conceptual errors due to the choice of an inconsistent case, or numerical approximation errors due to the choice of implementation scheme. Hence, the degree of symmetry in the calculated MCM can be understood as an indicator of “trustworthyness.”

We summarize that the only safe choice seems to be Case 4 because it is logically consistent, it handles real noisy data, and it is numerically straightforward to implement. Scientifically, we would be more interested in obtaining results for Case 1. This raises two questions: (a) Is the difference in results for the four cases actually significant in practical applications? And, if yes, (b) can we find a reasonably good approximation via any of the other cases to the desired Case 1? We will investigate these questions in three analytical scenarios (Section 4) and a real-world hydrogeological case study (Section 5).

4. Didactic Illustration With Analytical Scenarios

To illustrate how the four ways to consider or ignore measurement noise affect the outcome of BMS, we first design three didactic examples. We choose these scenarios such that (a) the model output distribution is one-dimensional and therefore easy to visualize, (b) BME values can be obtained analytically and therefore all four cases can be determined accurately, and (c) typical challenges encountered in real-world studies are represented such as models of varying predictive uncertainty, bounded support of the output PDF, and the search for an appropriate surrogate model.

The setup and motivation of the three scenarios is explained in Section 4.1, and results (posterior model weights and model confusion weights) are presented in Section 4.2. In Cases 2 and 4 (noisy data), posterior model weights are reported as averaged values: the expected value over \( N_d = 1,000 \) possible realizations of noise \( \epsilon \) is numerically approximated by Monte Carlo sampling as described in Section 3.4.

The source code of the analytical scenarios is available at https://bitbucket.org/Reuschen/measurement-noise-demo.

4.1. Setup of Scenarios

4.1.1. Scenario 1: Predictive PDFs of Varied Location and Width

In the first analytical example, we consider three models as three Gaussian predictive PDFs of varied location and width as defined in Table 1. They are visualized in Figure 4a. These PDFs shall represent model output simulating physics only (Cases 1 and 2). A real-world interpretation could be models of varying

| Model | Distribution | \( \mu \) | \( \sigma \) |
|-------|--------------|-------|-----|
| Blue  | Normal       | 0     | 0.2 |
| Red   | Normal       | 3     | 1   |
| Yellow| Normal       | -2    | 2   |
| Noise | Normal       | 0     | 0.3 |

| Table 1 | Parameter Choices for PDFs in Scenario 1 (Predictive PDFs of Varied Location and Width) |
complexity and predictive skill that differ in their maximum a posteriori prediction of the true system and in their predictive uncertainty. Panel (b) shows the same three PDFs after convolution with the Gaussian measurement noise PDF and hence representing model predictions in Cases 3 and 4. The chosen error standard deviation $\sigma$ is also listed in Table 1.

4.1.2. Scenario 2: Predictive PDFs With Bounded or Semi-Infinite Support

The second analytical scenario features a normal, exponential, and uniform PDF to represent model output (Figure 5a). PDF parameters are given in Table 2. The focus is here on the effect of stark contrasts in the support of the competing models/PDFs: the uniform PDF has bounded support and hence assigns values outside of the supported range a probability of zero, and the exponential distribution has semi-infinite support of only non-negative values. These hard constraints should be honored by the model selection result. This didactic scenario is hence designed to illustrate the effect of choosing one of the four ways to account
for noise in the presence of bounded or semi-infinite PDF support (data ranges where at least one of the competing PDFs is zero).

This scenario is highly relevant in real-world applications since many variables of natural systems are affected by processes that lead to such hard constraints (e.g., conservation of mass, monotonous increase in entropy, water flowing downhill only, non-negative concentration values, saturation falling between 0 and 1). Real data affected by measurement noise might even lie outside of the value range supported by the model (if only predicting physics), but such data points would not automatically reject the model’s underlying hypotheses. Hence, in such realistic situations, it is even more important to carefully dissect whether models and data contain noise or not.

### 4.1.3. Scenario 3: Identification of a Surrogate Model

To complete the set of analytical scenarios, we wish to mimic the realistic case of identifying a simpler model as a suitable surrogate for a complex reference model. That means the model selection task is to choose the surrogate model that approximates the high-fidelity model best. This can be achieved using the MCM as proposed by Schäfer Rodrigues Silva et al. (2020). The PDFs considered in this scenario are defined by the parameters listed in Table 3, and are visualized in Figure 6a. We declare the blue model to be the high-fidelity model which should be approximated by either the red or yellow surrogate model.

### 4.2. Results and Discussion

For each of the three didactic scenarios, we first present posterior model weights as a function of measured data for the four different cases (Section 4.2.1), and then the full MCM (Section 4.2.2).

#### 4.2.1. Posterior Model Weights

**4.2.1.1. Scenario 1: Predictive PDFs of Varied Location and Width**

Figure 4a shows the predictive PDFs for the true system state obtained by the three competing models; Figure 4b shows the predictive PDFs modeling the full data-generating process (including noise), which are therefore wider than the pure-physics PDFs. In those subplots, the BME value of model $k$ corresponds to the height of the PDF of model $k$. For example, the blue model without noise in Figure 4a scores a BME value of 2 for $d_0 = 0$. Note that the data vector $d_0$ is a scalar value in all three analytical scenarios, since we are investigating one-dimensional predictive distributions for simpler illustration.

Further, Figures 4c–4f show the posterior model weights obtained in the four different cases of how to consider noise. In these plots, the $x$-axis corresponds to a noise-free data value $d_0$, and the $y$-axis corresponds to the posterior model weight a model achieves given this data value. In Cases 2 and 4, these are averaged values, because we randomly sample the measurement error to be added to the noise-free data and then average over the respective posterior model weights to find the corresponding $y$-value for a given data value $d_0$. Hence, the plots are to be read as follows: the height of each model curve represents the average posterior model weight, given that the true system state equals $d_0$.

Apparantly, the four cases yield different results. Let us investigate the data point $d_0 = 1$ in Figure 4 in more detail. The posterior model weights in Case 1 of models blue, red, and yellow are 0, 0.45, and 0.55, respectively. In Case 2, they are 0.05, 0.45 and 0.50; in Case 3, 0.16, 0.40 and 0.44, respectively. Finally, in Case 4, they are 0.26, 0.40, and 0.34. This means that, for a noise-free data value of $d_0 = 1$, the posterior weights per model vary between 0 to 0.26, 0.4 to 0.45, and 0.34 to 0.55, respectively, depending on the choice of how to treat noise. Especially the jump from 0 to 0.26 is dramatic in terms of Bayes factors and their linguistic interpretation (Jeffreys, 1961), where evidence is in favor of the best versus the worst
model would jump from “decisive” to “not worth more than a bare mention.” Importantly, not only the decisiveness in model selection (or rejection) changes but also the ranking itself: in Case 4, the red model wins, whereas in all other cases, the yellow one scores best. Hence, conclusions about model selection may change dramatically depending on the choice of the case to be in.

Further, we learn from Figure 4 that the differences between the four cases are smallest if one model clearly shows the highest predictive density and hence obtains a much higher BME value than the others (here, e.g., at the PDF modes $d_0 = -2, 0, 3$). Differences are highest in the “transition zones” where BME values are not significantly different and model ranking is less decisive. Unfortunately, these are often the scientifically most interesting situations.

4.2.1.2. Scenario 2: Predictive PDFs of Bounded or Semi-Infinite Support

Figure 5 visualizes the predictive PDFs of the three competing models in Scenario 2 without noise in panel (a) and containing noise in panel (b). In this scenario, the difference is much more pronounced and important than in scenario 1: here, adding noise to the models eliminates the bounds of model blue (originally only defined on the interval $[-1; 1]$, now a symmetrical PDF with infinite support) and the semi-infinite support of model yellow (now a skewed PDF with infinite support). Hence, if we consider noise in the models, results look similar to scenario 1 (in cases 3 and 4, panels (e) and (f)), with differences between the two cases of using noise-free or noisy data for model evaluation being relatively small. As expected, using noisy data dilutes the decisiveness of model ranking to some degree, which of course depends on the level of measurement error standard deviation; a flip in model preference is not observed in this specific example.

If, instead, we model physics only (Cases 1 and 2, panels (c) and (d)), it matters a lot whether we use noise-free or noisy data for BME evaluation. If we use noise-free data (Case 1), we directly evaluate the predictive PDFs in panel (a) to read off BME values, and hence, the blue model scores a model weight of almost one in its bounded interval because it shows a much higher predictive density than the red model and because the yellow model’s PDF is zero in this range. Outside of its interval, the blue model of course obtains a weight of zero. Similarly, the yellow model can only gain a weight larger than zero for non-zero values. Consequently, the red model scores a weight of one over those data value ranges where none of the other PDFs “lives,” simply because it is defined there, not because it shows a competitively high predictive density. This scenario has the potential to provide a lot of insight about the system under study: if we compared three competing hypotheses and were (magically) able to consult noise-free data of the true system state, we could clearly reject the individual hypotheses outside of their scope, and hence, only the one hypothesis survives that is general enough to cover the full data range.
If we use noisy data (Case 2), the decisiveness in model choice is practically smoothed out. This is due to the averaging over random samples of noise: we evaluate the predictive PDFs in panel (a) at several locations around a specific noise-free data value $\hat{d}_0$, and then average it to determine the posterior model weight. Averaging in a neighborhood of $\hat{d}_0$ is dangerous in this scenario, since we possibly cross the boundaries of the support of the individual model PDFs. This is why the sharp transitions of model preference as seen in panel (c) (Case 1) are diluted in panel (d) (Case 2). In reality, however, this averaging procedure cannot be done, because we only have one noisy observation $\hat{d}_0$. Consequently, we rely on determining $P(M_k|\hat{d}_0)$ instead of $\mathbb{E}_{\hat{d}}[P(M_k|\hat{d})]$.

**4.2.1.3. Scenario 3: Identification of a Surrogate Model**

In the surrogate scenario (Figure 6), we assume that the blue model is a high-fidelity model and the red and yellow models are possible surrogates, with the red model being an exact but shifted copy of the blue model (e.g., due to some simplification bias), and the yellow model being a much more uncertain description of the system with a bias in location. Under this setup, noise-free data is in fact available (by the high-fidelity model), so this is the only occasion where we can indeed freely choose between Case 1 (using noise-free data) and Case 4 (using noisy data).

The resulting weights are a mix of the results discussed for scenarios 1 and 2: only in Case 1, the bounded support of the blue and red model is honored in the posterior model weights. Cases 2, 3, and 4 show different types of smearing, with Case 2 being closest to Case 1 with respect to how far the high-weight region of each model extends (i.e., steepest slopes at the transition points), and Case 3 being closest to Case 1 with respect to decisiveness in model choice (i.e., highest weights for blue and red, but note that the high-weight region is shifted in comparison to Case 1). Section 4.2.2 presents how the MCM can be used to choose the best surrogate in scenario 3.

**4.2.1.4. Influence of Measurement Error Standard Deviation on Differences Between the Four Cases**

In general, the differences between Case 1 and Case 4 depend on the level of measurement noise. Here, we want to show how large the influence of measurement noise is. Figure 7 shows the posterior model weights obtained in Case 4 for different standard deviations of noise on the example of scenario 1 (PDFs of varied location and width). With a standard deviation of zero, all four cases are the same and collapse to Case 1 (noise-free model and data). Obviously, the difference between Case 4 and Case 1 increases with increasing measurement noise.

It might seem unlikely that the measurement error standard deviation exceeds the variability of the model prediction; however, in realistic cases, this might happen for strongly calibrated models that are underdispersive in their predictive distribution. To detect whether the present level of noise dominates the model selection result in a specific application, we recommend using the MCM as a diagnostic tool (Section 4.2.2).

**4.2.1.5. Approximations of Case 1**

Recall that Case 1 is scientifically most interesting but intractable in real-data scenarios due to the data-availability barrier depicted in Figure 3. Since large measurement noise leads to relevant differences between the four cases, we search for a suitable alternative to approximate the results of Case 1 as closely as possible with the right column of the decision matrix (Cases 2 and 4). Case 4 would be a preferable candidate because it is the only consistent case left; however, all three investigated scenarios have shown that results for Cases 1 and 4 significantly differ from each other. Especially scenario 2 teaches us to be very careful when analyzing Case 4 but interpreting it as a proxy for Case 1: there is a danger that we believe to have identified a plausible model although its core hypothesis about the functioning of the system (without considering noise) is plainly rejected (e.g., the blue model at $d_0 = -2$ in Figure 5).

![Figure 7. Posterior model weights as a function of measurement error standard deviation $\sigma_{\text{noise}}$ on the example of Case 4 scenario 1. For a standard deviation of zero, Case 4 collapses to Case 1.](image-url)
Rather, results suggest that, if we are interested in Case 1, we should use Case 2 to approximate Case 1. Although being logically inconsistent, posterior model weights produced under Case 2 generally show a higher agreement with those in Case 1 (Figures 4–6). Thinking of the decision matrix and its barriers (Figure 3), we cross one barrier (i.e., going from top-left to top-right, crossing the data-availability barrier) to find a good proxy for Case 1.

4.2.1.6. Sensitivity of Model Ranking Results to Actual Outcome of Noise

We showed that Case 2 can be used as a reasonable proxy for Case 1. When using this proxy, we should be aware that all Case 2 results are averaged results over random realizations of noise $\varepsilon$. In real-world applications we only have one realization of noise $\varepsilon_0$ and hence only one measurement $d_0 = d_0 + \varepsilon_0$. Naturally the next question arises: How large are the differences in Case 2 between the averaged results ($P(M_k|\hat{d}_0)$) and the results based on one realization of noise ($P(M_k|d_0)$)?

The largest difference lies in the fact that $P(M_k|\hat{d}_0)$ might be a overconfident posterior model weight if only one realization of data $\hat{d}_0$ is used. We illustrate this with a simple example. Assume $d_0 = -0.25$ in scenario 2 (Figure 5). The BMS analysis given this data value (Case 1) results in a posterior model weight of 1 for the red model and 0 for the other two (Figure 5c). Because the true system state is perturbed by random noise (Case 2), we will end up with one of three possibilities. First, $\varepsilon_0$ could be between $-0.25$ and 0.25, yielding $\hat{d}_0 = d_0 + \varepsilon$ between $-0.5$ and 0. For that range, Case 2 will report the same results as Case 1. However, this only happens with a probability of 38% according to the assumed distribution of measurement noise. Second, with 31% probability the measurement noise $\varepsilon_0$ is in the interval $[-1.25, -0.25]$ which leads to a noisy data value $\hat{d}_0$ in the interval $[-1.5, -0.5]$. For these data values, Case 2 will report a posterior model weight of almost 1 for the blue (not the red) model. Third, the measurement noise $\varepsilon_0$ could be in $[0.25, 1.25]$(31%), which results in the yellow model being most likely. Hence, depending on the actual outcome of noise $\varepsilon_0$, any of the three models could win the model ranking.

This uncertainty in which model scores highest is visualized in Figure 8. Instead of posterior model weights, we now show the probability of each model to win the model ranking (i.e., being the model with the highest posterior model weight) as a function of true system state $d_0$. We observe that, for noise-free data (left column), probabilities are binary (0 or 1), because being the best model is completely defined by the constellation of the predictive PDFs in the model set. For noisy data, the randomness in the actual outcome of noise $\varepsilon$ leads to probabilities of being the best model smaller than one in the transition areas between high-density regions of the three model PDFs. The uncertainty about which model is best is highest in Case 2 for data values between $-1.5$ and 1.5, as discussed above.

To summarize, the shown posterior model weights of blue/red/yellow in Case 2 (e.g., 0.3/0.45/0.25 in Figure 5b) are a weighted average of the described three extreme outcomes of model ranking. If only a single data set $\hat{d}_0$ is available, this average cannot be taken and only one of the three extreme cases will occur. Hence, in practical applications, we need to keep in mind that the obtained ranking might be overly decisive (see also a preliminary summary of findings in Section 4.2.3).

4.2.2. Model Confusion Weights

Building on the obtained posterior model weights, we now use the model confusion matrices (MCMs) of the four cases to further investigate the BMS results. We here exemplarily discuss the results of scenario 3. Figure 9 displays the MCM obtained for scenario 3. Panel (a) shows the MCM in Case 1, panel (b) shows the MCM in Case 2, panel (c) the MCM in Case 3, and panel (d) in Case 4. For the results of scenarios 1 and 2, we refer the reader to Figures A1 and A2 in Appendix A.
4.2.2.1. Self-Identification Potential

The MCM can show if less noisy measurements can lead to more decisive posterior model weights. To do so, the self-identification weights (diagonal entries) can be used to compare the (expected) self-identification of Case 1 and Case 4.

The self-identification potential of the (blue) physics alone (Case 1, $A=91\%$) is much higher than the self-identification potential of the data-generating process (Case 4, $A=67\%$). This is expected because measurement noise (when conceptually attached to the model predictions by convolution) smears out differences between the model PDFs and hence makes them more similar (i.e., there is more potential for confusion). Case 1 can therefore be understood as an upper limit to the self-identification potential in Case 4 with the level of noise approaching zero. Practically, this means that we can investigate whether (and to which degree) more precise measurements could help in model selection by comparing the self-identification weights of Cases 1 and 4.

4.2.2.2. Symmetry of the Model Confusion Matrix

As mentioned earlier, the MCM of the consistent cases (Case 1 and Case 4) should result in a symmetric MCM, whereas the inconsistent cases should not (at least not necessarily). Generally, we can trace back any asymmetry in the MCM to either numerical problems or to an inconsistent choice of case. Either way, the symmetry of the MCM gives us an indication of how much we can trust our BMS results. If the MCM is symmetric, we can assume, yet not guarantee (because errors might cancel out), that everything is all right. If the MCM is asymmetric, we know that either numerics or philosophy should be a source of headache. Either way, this should motivate us to change something in our approach and/or implementation.

From Figure 9, we see that symmetry almost holds true for this analytical example in Case 1 and Case 4; we only see very small asymmetry on the level of rounding errors in Case 4. This happens due to the numerical approximation of the expected value over possible realizations of noise $\epsilon$ by Monte Carlo sampling. However, this numerical asymmetry is much smaller than the theoretically expected asymmetry of the inconsistent cases 2 and 3 (Figures 9b and 9c).

In practice, we can use the asymmetry in the MCM of Case 2 as an indicator for the (lack of) quality in approximating posterior model weights in Case 1 with Case 2. The more symmetric Case 2 is, the better it approximates Case 1 and the smaller the approximation error by using noisy data instead of noise-free data.

4.2.2.3. Identification of Most Suitable Surrogate Model

The MCM can be used to determine the best suitable surrogate model (Schäfer Rodrigues Silva et al., 2020). As shown hereafter, the best suitable surrogate model can depend on the choice of case.

Let us evaluate the red and yellow models’ potential as a surrogate model for the blue high-fidelity reference model. Considering that the red model has no overlap with the blue model when modeling physics only (Case 1, cf. Figure 6a), one would reject its use as a surrogate altogether. This is what we see in the MCM for Case 1; there is no confusion between the two models (confusion weight of 0%), and hence the yellow model would be preferred as a surrogate with a (still small) confusion weight of 9%.

If we instead aim to model the data-generating process (Case 4) as represented by the blue model, the red model is better suited because if the blue model produced the data, the model weight of red (19%) is slightly larger than the model weight of yellow (14%) (Figure 9d).

4.2.3. Summary and Implications of Findings From Analytical Scenarios

Our investigations on the three analytical scenarios have clearly demonstrated that the theoretical differences laid out in Section 3 yield significantly different results, for both posterior model weights, and for
model confusion weights. Hence, we need to be aware of the assumptions underlying each case and make them match the specific modeling context.

The logically inconsistent cases generally lead to problems in interpretation: Using noisy data to test noise-free model predictions (Case 2) leads to smearing over potential bounds or jumps in the predictive PDFs, and hence can mask severe differences between the hypotheses about the modeled system. Predicting the data-generating process but testing the competing models with noise-free data (Case 3) can be seen as an incomplete assessment of Case 4, because only the expected value of noise (=zero) is tested, which neglects the fact that posterior model weights are a nonlinear function of noise.

The logically consistent Case 4 suffers from a lower degree of identifiability per model because model differences are smoothened by noise. Yet, for specific applications, this will be the right context to choose (e.g., predicting whether or not a certain threshold value will be exceeded in a measurement).

Case 1 is scientifically most interesting because it aims to compare the predictive model PDFs of the true system state, that is, it focuses on the way how the models explain the system under study. Unfortunately, it suffers from the data-availability barrier and can only be evaluated in the synthetic setup of the MCM, but not for model ranking given real observed data. We found that Case 2 is a better proxy for posterior model weights of Case 1 than Case 4, even though being logically inconsistent. The approximation error can then be estimated from the degree of asymmetry of the corresponding MCM.

One issue remains: model ranking may turn out overly decisive in favor of one model if only one measurement (or data set) is used, that is, if one cannot average over many repetitions of data noise as in most practical applications. Trusting this overly decisive model ranking is dangerous because it does not necessarily represent the ranking given the true system state.

5. Application to Real-World Hydrogeological Case Study

To demonstrate the impact of choosing between the four ways to account for measurement noise when comparing models to data under real-world conditions, we have implemented the model ranking and model confusion analysis for a hydrogeological case study. All scenarios discussed so far have featured illustrative, one-dimensional predictive PDFs. One-dimensional problems are beautiful, easy to visualize, and allow for analytical solutions. In real life, we typically have high-dimensional predictive distributions, which are expected to amplify the differences between the four cases that arose already in 1D. Further, we are not only facing the data-availability barrier, but also the numerical-approximation barrier, since analytical solutions typically do not hold in practical applications. Our hydrogeological case features 2D fully-saturated groundwater flow in a confined sandbox aquifer, with heterogeneity in hydraulic parameters being represented by four competing models. We first describe the case study setup in Section 5.1, then present and discuss results in Section 5.2.

5.1. Case Study Setup

5.1.1. Experimental Data

For our analysis, we will use the experimental data from Illman et al. (2010). They performed steady-state cross-hole pumping tests in a lab-scale sandbox aquifer. These drawdown data sets have been used to infer the spatial distribution of hydraulic conductivity in the sandbox via Bayesian updating in Illman et al. (2010) and Schöniger, Illman, et al. (2015).

The experimental setup is described in Illman et al. (2010). In the following, we will give a short summary. The sandbox is 1.93 m long, 0.826 m high and has a width of 0.0102 m. It is filled with natural sand layering as shown in Figure 10a.

48 horizontal hydraulic ports (“wells”) were installed to monitor the pressure loss or to perform pumping. The sandbox can be modeled as quasi-2D because the ports penetrate the entire width. In this study, we follow Schöniger, Illman, et al. (2015) and use the 36 monitoring ports that provided the largest signal-to-noise ratio during the experiment. We consider pumping in the six wells marked in blue in Figure 10a. For the model
selection task, we consider two data-availability scenarios: we first assume just a single pumping test has
been performed (and to obtain representative results, we average resulting model weights of the six data sets
corresponding to possible six well locations), and second we combine all six pumping tests into one data set.

5.1.2. Model Set

We here use the drawdown predictions of the four alternative groundwater models developed by Schöninger, Illman, et al. (2015). These models vary in their spatial representation of heterogeneity in hydraulic conductivity \( K \). All four models use \( Y = \ln(K(x)) \) as a random field model, but differ in their assumption on the spatial (correlation) structure of \( K(x) \): (a) homogeneous (effective) value, (b) zonation with a homogeneous value within each zone, using independent random variables for each zone, and the geometry of zones known from visual inspection of Figure 10a, (c) geostatistical interpolation between pilot points (e.g., RamaRao et al., 1995), and (d) fully geostatistical parameterization.

To get a visual impression of the models and their differences in the heterogeneity of \( K \), conditional realizations of each model are shown in Figures 10b–10e. These realizations show the best fit to drawdown induced by a single pumping test in port 44, marked as a black rectangle in each plot. For a detailed description of the models, we refer to Schöninger, Illman, et al. (2015).
5.1.3. Numerical Implementation of Model Ranking and Confusion Analysis

In this work, we investigate two different scenarios of data availability as mentioned above. In the first scenario, we only consider drawdown data induced by a single pumping test. This results in a simple numerical problem because only a few conditioning data are available, with the high-likelihood region being rather spread out. In the second scenario, we combine data from all six pumping tests, which leads to a much stronger conditioning effect and a very peaked likelihood surface to be sampled. This makes the approximation of BME and hence the computation of posterior model weights and model confusion weights numerically much harder.

We use $A_A = 1,000$ Monte Carlo realizations to sample the predictive distribution of the data-producing model. BME per model is approximated using $A_M = 200,000$ Monte Carlo realizations of the homogeneous model and $A_M = 1,000,000$ realizations for each of the other models.

Further, as discussed in Section 3.4, we approximate the “zero noise in models” of Case 1 and Case 2 with a “small noise in models” of $A_A = 0.001 [m]$. This is 4.6 times smaller than the actual level of measurement noise ($A_A = 0.0046 [m]$).

5.2. Results and Discussion

In the previous analytical scenarios, we investigated the question of what the posterior model weight would be if the noise-free data $d_0$ be equal to some value. We can visualize the answer in one dimension (one measurement). Here, we have 35 drawdown measurements (excluding the one at the pumping port) or 210 measurements (when combining data from all six pumping tests). As a result, we cannot visualize all possible $d_0$. We, therefore, focus on the posterior model weights for the noisy observed data $d_0$ (measured by Illman et al., 2010) and the model confusion weights in Figures 11 and 12.

5.2.1. Results Based on Individual Pumping Tests

The goal of BMS is to select the best model out of an ensemble of possible models. Here, we demonstrate how posterior model weights should be interpreted. To do so, let us first look at the (computationally simple) scenario with drawdown data from individual pumping tests in Figure 11. Given the data-availability barrier for real data, we can obviously only report posterior model weights given experimental data for Cases 2 and 4. Further, the posterior model weights in Case 2 ($P(M_1 | d_0)$, top right) and Case 4 ($P(M_4 | d_0)$, bottom right) are shown.

5.2.1.1. Which Model is Best (in Case 4)?

Given the data-availability barrier for real data, we can obviously only report posterior model weights given experimental data for Cases 2 and 4 in Figure 11. When inspecting those weights for Case 4, we find a 60% probability of the zonated model being true. This is more than the self-identification potential of the zonated model (48%) in the MCM. It might seem surprising to score a higher posterior model weight than the reference value in the MCM. However, there are two causes that can lead to such results: First, it can happen by chance. The MCM shows the expected self-identification weight (48%) and we can score a higher one (or
a lower one) just by coincidence. Second, we might have overestimated the level of measurement noise. Here, we chose the measurement noise standard deviation equal to the maximum observed measurement noise of 0.46 cm, despite the fact that lower noise levels were observed at many other ports (Schöniger, Illman, et al., 2015). We neglected the spatial distribution of noise in this study to reflect common simplified assumptions about measurement noise, but wish to remind the reader that the results of the BMS analysis are of course conditional to the choice of error description (and hence the choice of likelihood function).

5.2.1.2. Can Posterior Model Weights be Increased With Less Noisy Data?

We could increase our confidence in model selection by reducing the noise in measurements by more precise (expensive) experimentation. Figures 11a indicates that with measurement noise close to zero, a self-identification of 92% is possible.

5.2.1.3. Can We Trust the Results?

From the theoretical foundations laid out in Section 3.3, we expect that Cases 1 and 4 yield symmetric MCMs. Here, we find this confirmed for Case 4 but not for Case 1. This means that this MCM suffers from a numerical approximation error in either (a) sampling the synthetic data, in (b) evaluating BME, or in (c) approximating Case 1 with a small-but-not-zero noise level (cf. Section 3.4), and results should be interpreted with caution. These errors could be reduced by increasing the Monte Carlo ensemble size of both the BME approximation and random noise sampling; however, it is our intent to demonstrate that the level of asymmetry caused by potential numerical errors tends to be much smaller than the compete asymmetry seen in Cases 2 and 3 due to their logical inconsistency (potentially further aggravated by numerical errors).

5.2.1.4. Can We Use Case 2 as a Proxy for Case 1?

We could ask ourselves whether the posterior model weights for Case 2 could be used as an approximation to the ranking in Case 1 to identify true system behavior instead of the data-generating process. For our specific case study, we conclude that we should not interpret the weights from Case 2 in that direction, because the large asymmetry of the MCM of Case 2 reveals a large gap between the two cases.

5.2.1.5. How Can We Approximate the MCM of Case 1?

To determine the upper limit for self-identification we had a look at the MCM in Case 1. However, we may not be able to calculate the MCM in Case 1. As a result, we want to investigate which of the other cases is best in approximating the MCM of Case 1. Recall that we did not have to search for a proxy of the MCM in Case 1 in our analytical scenarios (Section 4) because BME could be calculated analytically.

Now, with the real-world example, we have to address the numerical approximation barrier. While Case 2 was found to be a reasonable proxy for posterior model weights, this does not hold for MCMs due to the role of variance in the synthetic data as explained below. Instead, we find that Case 3 yields, qualitatively, the most similar results to the MCM of Case 1. In fact, we approach Case 1 numerically by starting from Case 3 (with the noise level chosen a priori from the knowledge about the measurement process) and reducing it as much as numerically possible (here: standard deviation reduced by factor 4.6 to obtain the approximated result of Case 1).

5.2.1.6. Can We Explain the MCM Structure of the Inconsistent Cases?

From Figure 11, we observe that the MCM of Case 2 has higher values in the bottom left whereas the MCM of Case 3 has higher values in the top right. This does not happen by chance but because we sorted the models from low complexity (homogeneous) to high complexity (geostatistical). This observation shows that Case 2 overestimates the posterior model weight of complex models, whereas Case 3 underestimates the posterior model weight of complex models. This can be explained, again, with the variance of the models and the data: In Case 2, the variance in the data (columns) is much larger than in the tested models (rows), and hence, the model with the highest variance has the best chance of scoring non-zero likelihoods. The opposite is happening in Case 3: The data show less variance than the noise-free models, and hence, in the spirit of the bias-variance tradeoff implicitly performed by BMS (Schöniger, Illman, et al., 2015), the model with the smallest variance that still fits the data reasonably well will score the highest model weight.
5.2.2. Results Based on Six Pumping Tests

We will use this occasion to show that stronger data sets do not only increase numerical errors but make the inconsistent cases close to ridiculous.

5.2.2.1. Which model is best (in Case 4)?

Again, the data-availability barrier only allows us to report the posterior model weights given experimental data for Cases 2 and 4 in Figure 12. We find a 86% posterior model weight of the zonated model in Case 4. This weight is similar to the self-identification potential of the zonated model in the MCM (92%). Hence, we again are optimistic that the zonated model is the quasi-data-generating model and no "not-in-the-set model" generated the data (although the MCM cannot prove that, it could only reject that hope).

5.2.2.2. Can Posterior Model Weights be Increased With Less Noisy Data?

We find that reducing measurement noise will probably only slightly increase our confidence in choosing the zonated model because Case 1 only produces a slightly higher expected self-identification weight of 97%.

5.2.2.3. Can We Trust the Results?

The second (computationally much more challenging) scenario based on six pumping tests shows even more difficulties in the resulting MCMs (Figure 12). The matrix of Case 4 is almost symmetric \( P(M_4|M_3) \approx P(M_3|M_4) \) and hence we conclude that also the posterior model weights given experimental data can be trusted, especially because the MCM in the row and column of the largest posterior model weight (86%) is almost symmetric. Case 1, instead, is asymmetric. Hence, the results of Case 1 are corrupted by numerical errors. This means, that the upper limit of 97%, found previously, might not be trustworthy.

In Case 2, numerics completely broke down and all BME values for all models and all data sets are always zero. This happens because the random noise added to the 210 measurements produces completely unlikely data values as seen through the noise-free predictive distributions. As a consequence, the likelihood of any data set, which equals the multiplication of 210 independent likelihoods of single observations, was smaller than machine precision, and hence calculating average posterior model weights failed.

5.2.2.4. Can We Use Case 2 as a Proxy for Case 1?

The posterior weights of Case 2 should not be used for model selection due to the completely “broken” MCM. This “broken” MCM indicates that the inconsistency of Case 2 is enormous as discussed earlier. The resulting posterior weights cannot be trusted and they certainly should not be used as an approximation of Case 1.

5.2.3. Summary and Implications of Findings From Real-World Case Study

The results of our hydrogeological case study showed that the difference between the four cases is even more pronounced under real-world conditions with high-dimensional data sets and more dominant measurement noise. Further, posterior model weights of Case 2 can only be obtained in scenarios where we obtain BME analytically. In this case study (and mostly in practice), Case 2 is not available, so we recommend approaching posterior model weights of Case 1 from Case 4 by reducing the noise level in the likelihood function. To find out whether better (noise-free) data can make model choice more decisive, we recommend approaching the MCM of Case 1 from Case 3 by reducing the noise level in the likelihood function.

6. Summary and Conclusions

In this study, we discuss where and for which reasons measurement noise should be considered in Bayesian model selection. We distinguish between four different cases (accounting for noise in models and/or data: (a) no-no, (b) no-yes, (c) yes-no, (d) yes-yes that differ conceptually as visualized in Figure 2). We have demonstrated on three analytical scenarios and a real-world case study that these conceptual differences
result in significantly different outcomes of Bayesian model selection. Thus, knowing which case to use is of high practical relevance.

Cases 2 and 3 are logically inconsistent because noise is considered either in the models or in the data. Therefore, we focus on the two consistent cases 1 and 4. They answer the following research questions:

1. Case 1: Which model is best in modeling the physics?
2. Case 4: Which model is best in predicting the data-generating process (i.e., physics plus noise)?

Philosophically, the choice is clear: either you are interested in pure physics, then you choose Case 1, or you are interested in simulating the data-generating process, then you move to Case 4. We have shown that model selection results under those two modeling goals do not necessarily agree and that the differences are especially pronounced for

1. Large measurement noise
2. Large data sets
3. Sudden jumps in the predictive density of models, especially to zero probability

Practically, we face two challenges as visualized in Figure 3: First, in real-world applications, the data-availability barrier blocks us from using Case 1 and Case 3 for model selection because noise-free data is unavailable. One notable exception is the task of identifying a suitable surrogate model for an (expensive) complex model because this analysis does not involve real observed data. Second, if no analytical formulation of Bayesian model evidence (BME) is available, we can only numerically approximate Case 1 and Case 2 with bias (e.g., by using a narrow likelihood function instead of a Dirac delta function). We call this the numerical approximation barrier. As a result, only Case 4 can be evaluated straightforwardly.

To not give up on the scientifically more interesting Case 1, we investigated the potential of approximating Case 1 with any of the other three cases. From the analytical scenarios, we found that Case 2 is the best proxy of posterior model weights of Case 1. Hence, the inconsistent Case 2 is the best approximation of the "physics" case and a "default fall-back" to Case 4 is not optimal. Remember that Case 2 is only available if BME can be obtained analytically. In a general modeling context, Case 2, therefore, needs to be numerically approximated by reducing the noise level in the likelihood function of the model, effectively pushing the numerical-approximation barrier from Case 4 toward Case 2.

The model confusion matrix (MCM) can be used to evaluate how much we can trust the obtained posterior model weights. The MCM can be computed for all cases (with numerical challenges in Cases 1 and 2, as mentioned above). We have shown mathematically that the MCM of consistent cases is by design symmetric. Any asymmetry in the MCM, therefore, indicates that either numerical errors or logical inconsistencies exist. The degree of asymmetry can be understood as a warning of how much the (approximated) model selection results lack interpretability. Further, the MCM can be used to get insight on whether or not a better (almost) noise-free measurement device would make model selection more decisive: The self-identification potential of models in Case 1 provides an upper limit to decisiveness in model choice under minimal noise.

Based on our findings, we propose the following procedure for Bayesian model selection in the presence of measurement noise (visualized in Figure 13). First, the philosophical question (a) must be answered: are we interested in modeling physics or the data-generating process? Then, dependent on whether or not noise-free data is available (b), the posterior model weights of the corresponding case are calculated (c). Next, the corresponding MCM is calculated as a reference for interpreting the posterior model weight results (d). After that, we check the MCM for asymmetry, which reveals the degree of numerical errors and/or logical inconsistency (e). If the resulting MCM is strongly asymmetric, the posterior model weights and the MCM are deemed unreliable and one might be forced to change the research question. Finally, the MCM of Case 1 can be determined or approximated to find out whether better (noise-free) data can make a model choice more decisive (f).

With this recommended procedure, modelers are forced to reveal (and make themselves aware) of their motivation for model selection, and results are ensured to be as consistent as possible, given the practical limitations of real-world data availability and numerical implementation. We believe that understanding the four ways of how to treat measurement noise in Bayesian model selection is relevant to any modeling
endeavor in water resources research and beyond, where measurement noise is significant enough to spoil a perfectly clear model choice.

Appendix A: Additional Results of Analytical Scenarios

A1 Model Confusion Weights of Scenario 1

Figure A1 displays the MCM obtained for scenario 1. Panel (a) shows the MCM in Case 1, panel (b) shows the MCM in Case 2, panel (c) the MCM in Case 3 and panel (d) in Case 4. Figure A1 reveals that the self-identification probability of the red model is almost independent of the chosen case. Differences are more pronounced for the blue and yellow model, with self-identification weights between 74% and 89% and between 77% and 83%, respectively. Cases 1 and 4 are practically symmetric; Cases 2 and 3 show some asymmetry between yellow/blue and blue/yellow.

A2 Model Confusion Weights of Scenario 2

Figure A2 displays the MCM obtained for scenario 2. Panel (a) shows the MCM in Case 1, panel (b) shows the MCM in Case 2, panel (c) the MCM in Case 3, and panel (d) in Case 4. In Case 2 the red/blue (1%) and blue/red (36%) entries are completely different. This indicates that Case 2 is a bad approximation of Case 1 if we want to select between the red and blue model. And indeed, we can see...
in Figure 5 that Case 1 and Case 2 differ a lot in the transition areas where the red and blue models show high posterior model weights.

In contrast, we can investigate the red and yellow models. Here, the off-diagonal entries (\(A_41\%\) and \(A_46\%\)) are similar. Hence, we follow that Case 2 is a good approximation of Case 1 if we are interested in selecting between the red and yellow model. Again, Figure 5 confirms this because Case 1 and Case 2 have similar behavior in the area where the red and yellow models have high posterior model weights.

Data Availability Statement

The data, models, and ensembles of Schöniger, Wöhling, and Nowak (2015) were used for the real-world case study.

References

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian Theory* (Vol. 405). John Wiley & Sons.

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

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

Elsheikh, A. H., Wheeler, M. F., & Hoteit, I. (2014). Hybrid nested sampling algorithm for Bayesian model selection applied to inverse subsurface flow problems. *Journal of Computational Physics*, 258, 319–337. https://doi.org/10.1016/j.jcp.2013.10.001

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (1995). *Bayesian data analysis*. CRC Press.

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

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

Höge, M., Guthke, A., & Nowak, W. (2020). Bayesian model weighting: The many faces of model averaging. *Water*, 12(2), 309. https://doi.org/10.3390/w12020309

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, 54(3), 1688–1715. https://doi.org/10.1002/2017WR021902

Illman, W. A., Zhu, J., Craig, A. J., & Yin, D. (2010). Comparison of aquifer characterization approaches through steady state groundwater model validation: A controlled laboratory sandbox study. *Water Resources Research*, 46(4). https://doi.org/10.1029/2009WR007745

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

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

Liao, F., Di Baldassarre, G., & Montanari, A. (2009). Model selection techniques for the frequency analysis of hydrological extremes. *Water Resources Research*, 45(7). https://doi.org/10.1029/2007WR006666

Lartillot, N., & Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2), 195–207. https://doi.org/10.1080/1063515050043172

Leube, P., Geiges, A., & Nowak, W. (2012). Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. *Water Resources Research*, 48(2). https://doi.org/10.1029/2010WR010137
Liu, P., Elshall, A. S., Ye, M., Beerli, P., Zeng, X., Lu, D., & Tao, Y. (2016). Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. *Water Resources Research*, 52(2), 734–758. https://doi.org/10.1002/2014WR016718

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

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

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

Najafi, M., Moradkhani, H., & Jung, I. (2011). Assessing the uncertainties of hydrologic model selection in climate change impact studies. *Hydrological Processes*, 25(18), 2814–2826. https://doi.org/10.1002/hyp.6043

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, 61(9), 1666–1678. https://doi.org/10.1080/02626667.2016.1183009

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

Nowak, W., Rubin, Y., & de Barros, F. P. J. (2012). A hypothesis-driven approach to optimize field campaigns. *Water Resources Research*, 48(6). https://doi.org/10.1029/2011WR010161

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

Rama Rao, B. S., LaVenue, A. M., De Marsily, G., & Marietta, M. G. (1995). Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: I. Theory and computational experiments. *Water Resources Research*, 31(3), 475–493. https://doi.org/10.1029/94WR02258

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, e2020WR028100. https://doi.org/10.1029/2020WR028100

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

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

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

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4), 833–859. https://doi.org/10.1214/06-BA127

Volpi, E., Schoups, G., Firmani, G., & Vrugt, J. A. (2017). Sworn testimony of the model evidence: Gaussian mixture importance (game) sampling. *Water Resources Research*, 53(7), 6133–6158. https://doi.org/10.1002/2016WR020167

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1), 92–107. https://doi.org/10.1006/jmps.1999.1278

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., & Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2), 150–160. https://doi.org/10.1093/sysbio/syr085

Ye, M., Meyer, P. D., & Neuman, S. P. (2008). On model selection criteria in multimodel analysis. *Water Resources Research*, 44(3). https://doi.org/10.1029/2008WR006803