Computational Geosciences (2023) 27:45–62
https://doi.org/10.1007/s10596-022-10179-x

ORIGINAL PAPER

The method of forced probabilities: a computation trick for Bayesian model evidence

Ishani Banerjee[1] · Peter Walter[1] · Anneli Guthke[2] · Kevin G. Mumford[3] · Wolfgang Nowak[1]

Received: 8 July 2022 / Accepted: 18 October 2022 / Published online: 23 November 2022

Abstract

Bayesian model selection objectively ranks competing models by computing Bayesian Model Evidence (BME) against test data. BME is the likelihood of data to occur under each model, averaged over uncertain parameters. Computing BME can be problematic: exact analytical solutions require strong assumptions; mathematical approximations (information criteria) are often strongly biased; assumption-free numerical methods (like Monte Carlo) are computationally impossible if the data set is large, for example like high-resolution snapshots from experimental movies. To use BME as ranking criterion in such cases, we develop the "Method of Forced Probabilities (MFP)". MFP swaps the direction of evaluation: instead of comparing thousands of model runs on random model realizations with the observed movie snapshots, we force models to reproduce the data in each time step and record the individual probabilities of the model following these exact transitions. MFP is fast and accurate for models that fulfil the Markov property in time, paired with high-quality data sets that resolve all individual events. We demonstrate our approach on stochastic macro-invasion percolation models that simulate gas migration in porous media, and list additional examples of probable applications. The corresponding experimental movie was obtained from slow gas injection into water-saturated, homogeneous sand in a 25 x 25 x 1 cm acrylic glass tank. Despite the movie not always satisfying the high demands (resolving all individual events), we can apply MFP by suggesting a few workarounds. Results confirm that the proposed method can compute BME in previously unfeasible scenarios, facilitating a ranking among competing model versions for future model improvement.

Keywords Bayesian model selection · Markov-chain models · Gas-injection in porous-media · Invasion percolation models

1 Introduction

Many competing conceptual models can be used to represent real-world systems. These models differ in their underlying hypotheses, which need to be tested against real-world observation data for their accuracy in representing the featured real-world system. Bayesian model selection is a statistical method used for testing competing conceptual models against each other by ranking them based on Bayes' Theorem. Bayesian model selection involves computing Bayesian Model Evidence (BME), which is the likelihood of a model producing the observed data, given the prior distribution of its parameters.

Computing BME using analytical solutions is applicable only under strongly limiting assumptions, which generally do not hold in real-world applications [1]. So, other techniques involving mathematical approximations and numerical methods have been developed, but they all have their own limitations. Mathematical approximations (commonly known as Information Criteria (IC)) include the Kashyap information criterion (KIC) [2], the Bayesian information criterion (BIC) [3], the Akaike information criterion (AIC) [4] and so on. They are based on different assumptions and/or asymptotics. These criteria have been shown to yield misleading model ranking results in real applications if their assumptions are violated [5–13]. Using numerical methods [3] to compute BME avoids such assumptions but requires high computational effort. Numerical approximations that are commonly used for highly complex models are Monte Carlo (MC) methods with various sampling strategies, such as brute-force MC integration, MC integration with importance sampling, or MC integration with posterior sampling [1, 14, 15]. MC methods generally require large ensemble sizes and a good overlap between the likelihood function and the parameter

✉ Ishani Banerjee
ishani.banerjee@iws.uni-stuttgart.de

Extended author information available on the last page of the article.

prior. The latter corresponds to a well-specified prior and relatively uninformative data sets. This, in turn, means that practical applications can require extremely (up to prohibitively) large MC ensembles.

The size of the sampling ensemble for these methods is limited by the available computational resources. For high-dimensional problems (i.e. with many uncertain parameters), the so-called curse of dimensionality kicks in, requiring an exponential number of model evaluations [16, 17]. Additionally, for highly accurate or informative data sets, the overlap between predictive distributions and observed data may be so small that MC methods may not result in a meaningful BME value (> 0) at all.

For a model-data system involving binary (yes/no) decision output, the likelihood function becomes a Dirac-delta function, thus leading to likelihood values of zero for practically all sampled parameter values of the model. Thus, the BME value would tend to zero, and any model would be rejected as infinitely poor. This becomes a problem, especially for long-time sequences of repeated outputs. For example, in a lotto game, getting the first number right is not that difficult, but getting the exact sequence of six numbers in a row right is almost impossible.

For such model-data systems involving binary output, with highly discretized atomic-event-type data and Markov chain models, we propose a method to compute BME with a reasonably low computational effort. Observed states are called *atomic events* if each individual possible outcome can be enumerated and they are mutually exclusive and collectively exhaustive. *Markov Chain models* are stochastic models that fulfil the Markov Chain property, i.e. the probability distribution of model states in the next (time) step depends solely on the previous step, not on any prior state to that. We call our method of BME computation the *Method of Forced Probabilities* (MFP) due to its core idea: instead of evaluating millions of forward runs that may fit the data by random chance, the model is forced to follow the data during each time step. We record the individual probabilities of the model performing these exact transitions as if they were done without any constraints. Following a strict mathematical derivation, we compute BME as the product of these probabilities. By exploiting the Markov Chain property of the model with this procedure, we are able to compute BME in previously nearly impossible cases without resorting to any kind of approximations.

Model order reduction techniques offer an alternative approach for optimization, parameter sampling or Bayesian analysis of high-dimensional problems. Reduced-order models are a computationally cheap abstraction of the original, high-fidelity models [18]. Examples of such reduced-order modelling techniques include, but are not limited to, models obtained using projection-based model reduction method (e.g., polynomial chaos expansion [19], proper orthogonal decomposition [20]), response surface models (e.g., polynomials, kriging, radial basis functions, artificial neural networks, etc.[21]) and, lower-fidelity models (physically reliable simple abstractions of the system under study [21]). Although such reduced-order models assist in solving the computing time problem, they are only approximate. In contrast, our method (MFP) is exact. Also, our method tackles the challenge of evaluating BME rather than the computational efficiency issue of complex high-fidelity models. Further, our method can be used in combination with all reduced-order modelling approaches that maintain the Markov property. Other options include an abstraction of summary statistics from data (so-called approximate Bayesian computation [22]), manual-visible techniques (like moments matching [23]), or the use of plausible, non-Bayesian metrics [24].

In Section 2, we discuss the mathematical formulation of BME (Section 2.1), introduce our MFP approach for computing BME (Section 2.2), and illustrate it on a didactic example (Section 2.3). In Section 3, we introduce our test case for demonstration: we apply the method on a Stochastic Macro-Invasion Percolation (SIP) model that simulates multiphase flow in porous media (Section 3.1). The corresponding highly resolved data set was obtained from an experiment with slow gas injection into water-saturated, homogeneous sand in an acrylic glass cell (Section 3.2). We also design a synthetic data scenario for the proof-of-concept of our method (Section 3.3), and we list the implementation steps of the MFP for the SIP model under the different data scenarios (Section 3.4). Further, we add a list of general algorithmic steps of MFP in Section 3.5. In a previous study [24], we used the (Diffused) Jaccard coefficient to facilitate a quantitative comparison of an invasion percolation model to the experimental data set used in this study. This technique only works on image-type data and is not free from information losses. Therefore, our proposal MFP is the first method ever that facilitates a fully Bayesian assessment of the SIP model and is free of information loss. Section 4 discusses the results obtained from the synthetic (Section 4.1) and real-data scenarios (Section 4.2). Finally, we summarize the contributions of this study, draw conclusions, provide an outlook towards future work, and list a few examples of applications where our method can be applied in Section 5.

2 Bayesian model evidence and its evaluation via the proposed method of forced probabilities

We present the concept and mathematical formulation of Bayesian Model Evidence (BME) in Section 2.1. Then, we introduce the concept of our approach to computing BME

(MFP) in Section 2.2. We illustrate our proposed method with the help of a didactic example in Section 2.3.

2.1 Bayesian model evidence

For N_m competing models M_k, $k = 1...N_m$ and observation data $\mathbf{y}_0$, the BME value BME_k of any model M_k can be evaluated as (Bayesian integral from [25]):

$$\begin{aligned} BME_k &= p\left(\mathbf{y}_0 \mid M_k\right) \\ &= \int_{U_k} p\left(\mathbf{y}_0 \mid \mathbf{u}_k, M_k\right) \cdot p\left(\mathbf{u}_k \mid M_k\right) d\mathbf{u}_k \\ &\equiv I_k, \end{aligned} \tag{1}$$

where U_k denotes the model's parameter space, $\mathbf{u}_k$ represents a random parameter vector with prior distribution $p(\mathbf{u}_k \mid M_k)$, and $p\left(\mathbf{y}_0 \mid \mathbf{u}_k, M_k\right)$ is the probability or likelihood of the parameter set $\mathbf{u}_k$ of the model M_k to have generated the observed data set $\mathbf{y}_0$.

The integral I_k over the entire parameter space is computationally expensive and can become infeasible (with no meaningful BME value) in the cases discussed in Section 1. A review of existing methods to determine BME can be found in [1]. To facilitate the introduction of our proposed method, MFP, we present here the approach of simple (or: brute-force) MC integration [26] of Eq. 1. The integrand is evaluated at randomly chosen points ($\mathbf{u}_{k,r}$) of the parameter space U_k, which are drawn from their prior distribution $p(\mathbf{u}_k \mid M_k)$. The mean of the evaluated likelihoods ($p\left(\mathbf{y}_0 \mid \mathbf{u}_k, M_k\right)$) provides the approximate value of the integral (referred to as $\hat{I}_k$):

$$\hat{I}_k = \frac{1}{N} \sum_{r=1}^{N} p\left(\mathbf{y}_0 \mid M_k, \mathbf{u}_{k,r}\right) \approx I_k, \tag{2}$$

with N being the number of MC realizations (ensemble size). To re-stress the problem: in applications with many precise data, the summands in this equation are (close to) zero with a probability very close to one, such that convergence can be prohibitively slow.

To rank models against each other, one can directly compare their BME values (the larger, the better) or their negative logarithmic BME values (the smaller, the better). Alternatively, one computes so-called Bayes factors (BF) [25] for two models $k1$ and $k2$:

$$\mathrm{BF}_{\frac{k2}{k1}} = \frac{BME_{k2}}{BME_{k1}}, \tag{3}$$

with a scale for interpretation provided by, e.g., [27].

2.2 Method of forced probabilities (MFP): key idea

For our purposes, we redefine Eq. 1 as:

$$I_k = \iint_{U_k} p\left(\mathbf{y}_0 \mid \boldsymbol{\omega}_k, \boldsymbol{\theta}_k, M_k\right) \cdot p\left(\boldsymbol{\omega}_k, \boldsymbol{\theta}_k \mid M_k\right) d\boldsymbol{\omega}_k \, d\boldsymbol{\theta}_k. \tag{4}$$

Here, the parameter space U_k is split into uncertain parameters $\boldsymbol{\theta}_k$ and random events $\boldsymbol{\omega}_k$, $p(\mathbf{y}_0 \mid \boldsymbol{\omega}_k, \boldsymbol{\theta}_k, M_k)$ is the likelihood of the parameters $\boldsymbol{\omega}_k$ and $\boldsymbol{\theta}_k$ of model M_k to have generated the data set $\mathbf{y}_0$, and $p(\boldsymbol{\omega}_k, \boldsymbol{\theta}_k \mid M_k)$ is the prior probability density of these parameters. Uncertain parameters $\boldsymbol{\theta}_k$ comprise those parameters and inputs of the model with unknown or non-measurable values. Random events within the model $\boldsymbol{\omega}_k$ represent apparently stochastic system behaviour that cannot be explained deterministically (but only distribution-wise) by the model's equations, assumptions or mechanisms (see also Section 3.1).

Using the law of total probability [28], we split the double integral of Eq. 4 into an inner integral over random events and an outer integral over uncertain parameters:

$$\begin{aligned} I_k &= \int \left[\int p\left(\mathbf{y}_0 \mid \boldsymbol{\omega}_k, \boldsymbol{\theta}_k, M_k\right) \cdot p\left(\boldsymbol{\omega}_k \mid \boldsymbol{\theta}_k, M_k\right) d\boldsymbol{\omega}_k \right] \\ &\quad \cdot p\left(\boldsymbol{\theta}_k \mid M_k\right) d\boldsymbol{\theta}_k \\ &= \int p\left(\mathbf{y}_0 \mid \boldsymbol{\theta}_k, M_k\right) \cdot p\left(\boldsymbol{\theta}_k \mid M_k\right) d\boldsymbol{\theta}_k. \end{aligned} \tag{5}$$

The key idea of the Method of Forced Probabilities is to replace the inner integral (over random events) with a single analytical solution and use an MC integration (Eq. 2) only for solving the outer integral over uncertain parameters ($\boldsymbol{\theta}_k$), for models obeying the Markov Chain property. This means that for random events $\boldsymbol{\omega}_k$, as opposed to simulating thousands of forward model runs and waiting for a random match with the observed data, we instead record the individual probabilities $p(\boldsymbol{\omega}_k \mid \boldsymbol{\theta}_k, M_k)$ of the model performing the exact transitions observed in the data at each time step. Using the Markov chain property, the product of these probabilities corresponds to $p(\mathbf{y}_0 \mid \boldsymbol{\theta}_k, M_k)$ in Eq. 5:

$$p(\mathbf{y}_0 \mid \boldsymbol{\theta}_k, M_k) = \prod_{t=0}^{t_{max}-1} \mathrm{P}\left(\mathbf{y}_0(t+1) \mid \mathbf{y}_0(t), \boldsymbol{\theta}_k, M_k\right), \tag{6}$$

where $\mathrm{P}\left(\mathbf{y}_0(t+1) \mid \mathbf{y}_0(t), \boldsymbol{\theta}_k, M_k\right)$ is the probability of transition in $\mathbf{y}_0$ (in accordance with the data) from time step t to $t+1$, and t_{max} is the total number of time steps in the experimental data. The idea is to plug this exact analytical solution into Eq. 5 and use the MC method only for the uncertain parameters.

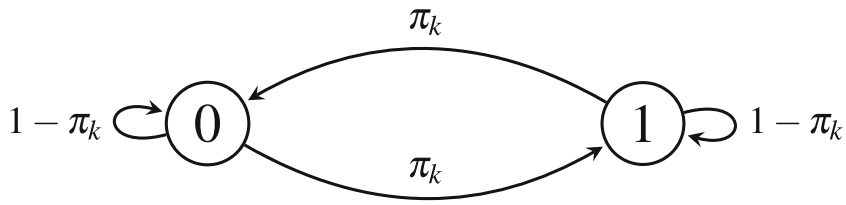

Fig. 1 Transition graph of toy Markov chain model

If numerical scaling becomes an issue for Eq. 6, one can simply work in (negative) logarithmic scale:

$$
\begin{aligned}
&-\ln p(\mathbf{y}_0 \mid \boldsymbol{\theta}_k, M_k) \\
&= -\sum_{t=0}^{t_{max}-1} \ln P\left(\mathbf{y}_0(t+1) \mid \mathbf{y}_0(t), \boldsymbol{\theta}_k, M_k\right). \qquad (7)
\end{aligned}
$$

Further, even after using the logarithmic scale, numerical issues with the BME values can arise during averaging (after exponentiating Eq. 7) for the outer integral of Eq. 5 due to the scale and span of individual values. We address this by a numerical trick that involves subtracting a common BME value at the logarithmic scale, such that the exponent of Eq. 7 (Eq. 5) is closer to zero, see Appendix B.

One may argue that the act of multiplying individual likelihoods in order of appearance in a time sequence is close to the process done in data assimilation methods, where a time series of time slice-wise likelihoods and cumulative BME values can be spit out as a simple by-product. This analogy is most apparent when comparing to particle-filter-like schemes for data assimilation [29]. Moreover, just like our BME computation can be used for parameter selection / Bayesian update of parameters, this could also offer the path to parameter estimation in data-assimilation mode. Some data assimilation schemes perform a joint estimation of system states *and* uncertain parameters, typically called *augmented state vector* approaches (e.g., [30]) or *parameter-space* schemes (e.g., [31, 32]). Without going into further detail, this opens a future pathway to apply our MFP method in (real-time) data assimilation for either state forecasting, parameter updating, or both at once.

2.3 MFP: implementation illustrated with a didactic example

As a toy model for demonstrating our method, let us consider a simple Markov Chain with two output states (0 and 1) and a fixed (instead of uncertain) parameter π_k (e.g., a repeated coin flip experiment). Thus, the Bayesian integral in Eq. 4 simplifies to:

$$
\begin{aligned}
&p\left(\mathbf{y}_0 \mid M_k\right) \\
&= \int_{U_k} p(\mathbf{y}_0 \mid \boldsymbol{\omega}_k, \boldsymbol{\theta}_k, M_k) \cdot p(\boldsymbol{\omega}_k \mid \boldsymbol{\theta}_k, M_k) d\boldsymbol{\omega}_k, \qquad (8)
\end{aligned}
$$

i.e. $\boldsymbol{\theta}_k$ is fixed, and the outer integral disappears. Note here that U_k only contains the random events. The transition probabilities of the model are defined as:

$$
\mathbf{P}_k(b \mid a) = \begin{cases} \pi_k & \text{if } b \neq a \\ 1-\pi_k & \text{if } b = a \end{cases} \qquad (9)
$$

Here, b is an output state at a particular flip, and a is an output state in the previous flip (see Fig. 1).

For a number of flips $t_{max} = 3$ (i.e., $t = 0, 1, 2, 3$), the possible predictions by the model are shown in the

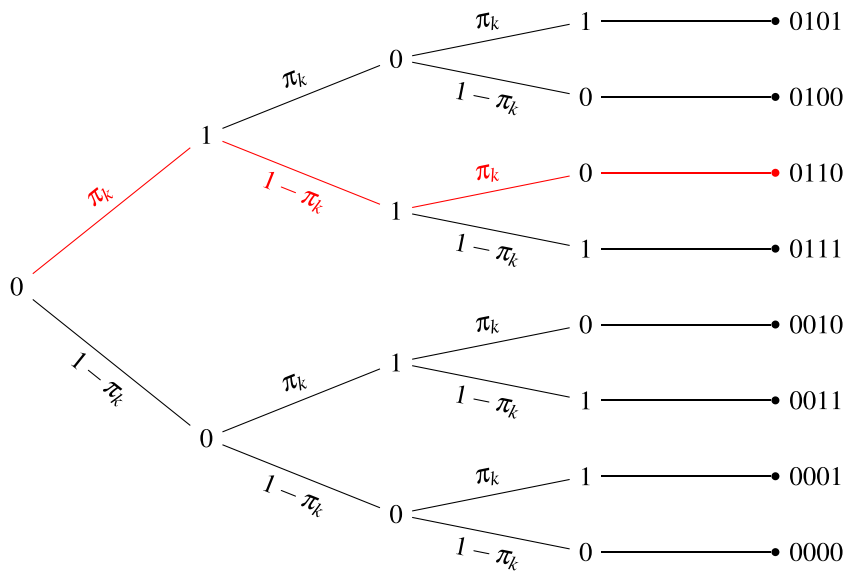

Fig. 2 Probability tree diagram for the toy Markov chain model with $t_{max} = 3$. The true sequence or observed series of outcomes is highlighted in red

probability tree diagram in Fig. 2. Additionally, in this diagram we fix the initial condition at $t = 0$ to $y(0) = 0$. Let us assume that the true observation data sequence is 0110 (highlighted in red in Fig. 2).

In such a simple tree structure with equiprobable branching ($\pi_k = 0.5$), it is obvious that the probability (BME) of observing the single true path with likelihood one is $\frac{1}{\text{number of paths}}$. Now imagine if the sequences' length t_{max} increases (a deeper tree) or the dimension of the state space is increased (more than two branches for each node), the complexity of the probability tree diagram will increase exponentially (see Appendix A for more details). For example, for a binary tree with $t_{max} = 100$, we end up with 2^{100} different paths. This would further diminish the BME value and increase the computational effort to completely sample all possible paths in direct MC-based approaches based on the conventional Eq. 2.

Most real-world applications involve a more complex structure, where the branches are not equiprobable or complete enumeration is not possible anymore. In such cases, an MC approach would be used to sample each random path in proportion to its probability, requiring an even more significant number of samples to represent all paths, including the ones with very low probability, statistically sufficiently well. Note that it is not enough to "hit" the one path that coincides with the observation, but for an accurate approximation of BME, we need an accurate representation of low probabilities just as well (zeros play an essential role in arithmetic averaging), see, e.g. [1].

In contrast, the MFP simply calculates BME as the probability of mimicking the observed state changes, i.e., a flip from 0 to 1 and then staying at 1 and finally flip again to 0:

$$p(\mathbf{y}_0 \mid \boldsymbol{\theta}_k, M_k) = \pi_k \cdot (1 - \pi_k) \cdot \pi_k.$$

With $\pi_k = 0.5$, this equals to the enumeration or MC solution of $\frac{1}{8}$. This means that we only need to calculate a finite product over a set of t_{max} (here: three) values. Therefore, our method (MFP) scales linearly with t_{max} and does not exponentially explode like full enumeration or MC methods.

3 Demonstration on a real case study

In this section, we demonstrate the applicability of our method on a more complex model with Markov Chain property: a Stochastic Macro-Invasion Percolation model (SIP), see Section 3.1. Invasion percolation (IP) models are discrete growth models, which repeatedly apply so-called rules that specify why which model block is invaded by the wetting or non-wetting fluid in the next step, see Section 3.1. Macro-Invasion Percolation (Macro-IP) models are an upscaled abstraction of pore-scale IP models. We have used a Macro-IP model with the experimental data used in this study in a previous study [24]. The SIP model is conceptually similar to the Macro-IP model of [24]. The difference between the SIP model and the Macro-IP model of [24] is an additional rule for stochastic selection from the Stochastic Selection and Invasion (SSI) model of [33]. Thus, to describe the SIP model, we describe the model formulation of the Macro-IP model [24] and the additional rule for stochastic selection in Section 3.1. We compare the SIP model to experimental binary-image data from gas-injection in homogeneous, water-saturated sand [34], see Section 3.2. For a test in the absence of all problems that real experimental data bring about, we also use synthetic model-generated data, see Section 3.3. Then, we discuss the implementation of MFP on the SIP model in Section 3.4 for both the synthetic and the real data sets. Before discussing the results of this case study, in Section 3.5, we list the general algorithmic steps of MFP.

3.1 Stochastic macro-invasion percolation model (SIP)

Originally derived from the Percolation theory of [35], Invasion Percolation (IP) models are often used for simulating multiphase flow in porous media (E.g., [23, 33, 36–46]). Various versions of IP models have been used in literature, but all of them have a similar implementation structure (illustrated in Fig. 3):

- The porous medium is conceptualized as a network of 2-dimensional (2D) or 3-dimensional (3D) blocks or nodes, with a given connectivity, by assigning threshold values to the blocks from a distribution (depending on the specific porous medium).
- Initially, all the blocks are occupied by one (defending) fluid. Then the invading fluid is placed in one of the blocks, depending on its source or injection site (see Fig. 3**a**).
- The neighbouring blocks of the invading fluid block are evaluated for their entry thresholds (pressure) and based on some rules (mostly minimum entry threshold, see below) one of the connecting blocks is filled by the invading fluid (see Fig. 3**b** & **c**). The filling process of one block can occur in a single step or in multiple steps.

Macro-IP models differ from traditional IP models in scale, i.e. the blocks in Macro-IP models represent a sub-network of pores and throats in contrast to individual pores as in IP models [24]. We use a 2D representation of these models to mimic the quasi-2D setup of the experimental data (Section 3.2) and then to simulate gas invasion in homogeneous water-saturated sand.

In the Macro-IP model [24], at each time step, the gas invades one of its neighbouring water-filled blocks using

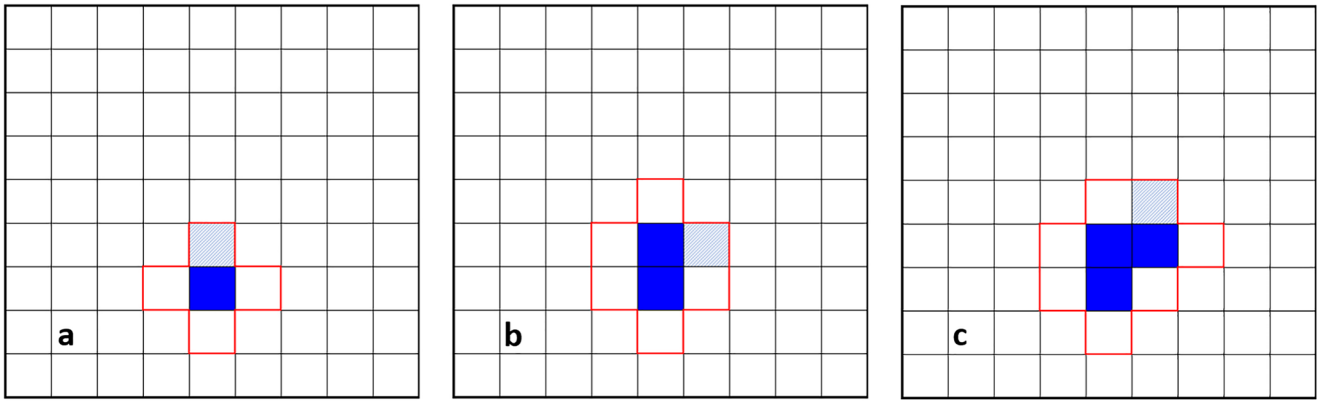

Fig. 3 Schematic to illustrate a general 2D Invasion Percolation (IP) model. Blue blocks are blocks filled with the invading fluid, and all other blocks are filled with the defending fluid. The hatched light blue block shows the next block to be filled and, the red-rimmed neighbouring blocks are the ones evaluated for invasion at each step

a rule that searches the block with the smallest invasion threshold T_e calculated using:

$$T_e = P_e + P_w, \tag{10}$$

where P_e is the local entry pressure in each block, which is the capillary pressure (P_c) required by gas to invade a water-occupied block. P_w is the pressure of the water phase, calculated with the hydrostatic pressure assumption as:

$$P_w = \rho_w g z, \tag{11}$$

where ρ_w is the water density, g is the acceleration due to gravity, and z is the height from the top of the acrylic glass cell. P_e is calculated from the Brooks-Corey capillary pressure (P_c) - saturation (S) curve [47]:

$$S_e = \frac{S_w - S_r}{1 - S_r} = \left(\frac{P_c}{P_d}\right)^{-\lambda}, \tag{12}$$

where S_e is the *effective* wetting-phase saturation, S_w is the wetting-phase saturation, S_r is the residual wetting saturation, P_c is the capillary pressure, P_d is the macroscopic displacement pressure, and λ is the pore-size distribution index, the value of which typically ranges between $1-4$ and can be up to 7 for very uniform sands. Please note that P_e is a specific value of P_c (a point on the $P_c - S$ curve), and that we provide the term: $\frac{S_w - S_r}{1 - S_r}$ of the Eq. 12 only to correspond with the general form of the Brooks-Corey pressure-saturation relation prevalent in literature. The term $\frac{S_w - S_r}{1 - S_r}$ has no further use in our model description.

At the scale of the model, the exact pore-scale arrangement is unknown. Thus, the capillary pressure P_c is randomized per block using the Inverse transform sampling method, the details of which have been discussed in [24]:

$$P_c = P_d \mathcal{U}^{-\frac{1}{\lambda}}, \tag{13}$$

with $\mathcal{U}$ being a random number from a standard uniform distribution on the interval [0, 1].

Since the Macro-IP model from [24] is deterministic for any given value of $\boldsymbol{\theta}$ and a frozen set of random P_c values, we would have no random events $\boldsymbol{\omega}$ within the model. That means computing BME would focus only on the outer parameter-related integral of Eq. 5. The inner integral would degenerate to a simple yes or no problem. Without addressing measurement errors or any other form of randomness between the model and data, the answer would be a straightforward rejection with BME $= 0$. Thus, to include this model in our BME comparison, a modification of the model to include random events $\boldsymbol{\omega}$ is required. This is achieved by using the SIP model.

SIP differs from the Macro-IP in the way that instead of gas selecting the block with the minimum invasion threshold (T_e) for invasion, it invades based on a slightly modified rule for stochastic selection from the Stochastical Selection and Invasion (SSI) model of [33]. The stochastic selection rule of the SSI model accounted for viscous effects (randomness brought into the system by the fluids) and was originally applied to dense non-aqueous phase liquid (DNAPL) migration. At the injection rate of the experimental data used in this study, viscous forces are negligible. However, viscous forces come into play at other injection regimes or even for different fluids. This stochastic selection rule has been modified to be applicable for our gas invasion in water-saturated sand [23], instead of the DNAPL invasion of the original work, and is explained below.

In the modified stochastic selection rule of the SSI model, the decision of gas invasion is still proportional to the T_e values of the neighbouring blocks but is slightly modified using an additional parameter: c called the cell selection weighting factor [33]. In this rule, the list of T_e values of

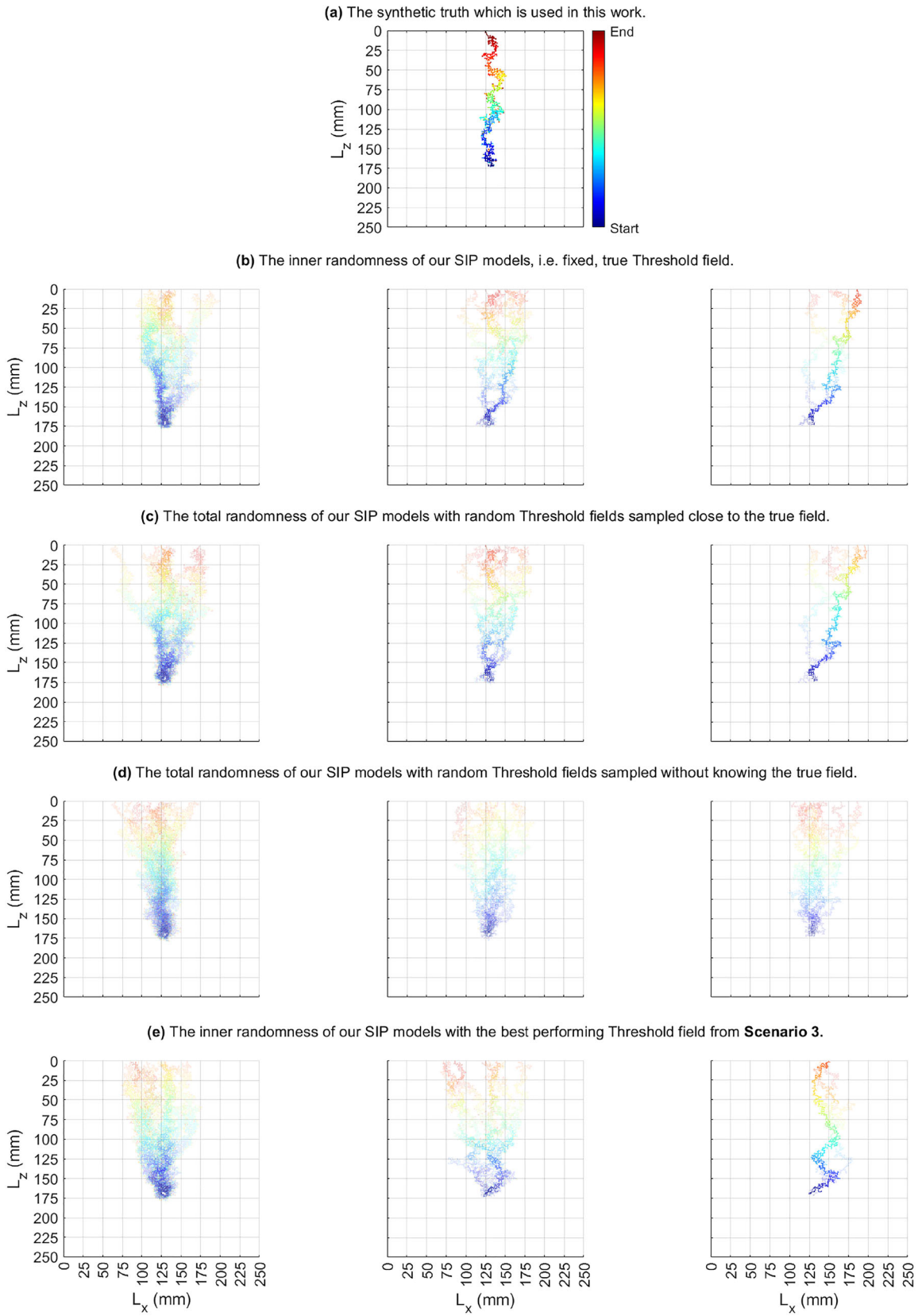

Fig. 4 To visualize the different SIP model versions and the synthetic scenarios (Section 3.4.1): this figure shows a sample of 10 model runs for each combination and compares them to the synthetic data set. More frequently invaded cells appear more opaque, and their colour shading indicates the relative time from the first to the last invaded cell. From left to right c values are 5, 15, and 100

the neighbouring blocks (n) of the current gas cluster are arranged in an ascending order $T_{e,asc}$ and the cumulative sum $T_{e,cum}$ is evaluated:

$$T_{e,cum}[i] = \sum_{j=1}^{j=i} T_{e,asc}[j]; i = 1, 2, 3, \ldots, n \tag{14}$$

Then, the first block (value of i) where the rule specified in Eq. 15 is found true is invaded by the gas:

$$T_{e,cum}[i] > \mathcal{R}^c \times \sum_{j=1}^{j=n} T_e[j] \tag{15}$$

Here, $\mathcal{R}$ is a uniformly distributed random number between [0,1], and c, in the range of [0....∞], controls the strength of randomness in the stochastic selection rule. When $c \to \infty$, the value of $\mathcal{R}^c \to 0$ for almost all values of $\mathcal{R}$. In this case, the block with the lowest T_e value, first on the list of $T_{e,asc}$, will be selected for invasion. This results in lightning-bolt-like gas fingers. The lower the c value, the higher the RHS of Eq. 15, which ensures that the higher $T_e[j]$ are picked more often; hence we observe a gas finger that is not moving strictly upward, but resulting in a gas finger pattern with a wider spatial distribution. Example simulations for $c = 5, 15, 100$ will be provided in Section 3.3, see, e.g., Fig. 4. Once the gas invades a block, it is assigned a gas-saturation value of 1. Thus, the model has a binary (gas or no gas) type of image data as output.

Furthermore, in our SIP model, we also include the additional re-invasion rule of the Macro-IP model from [24] to incorporate fragmentation and mobilization events [23, 41, 48] observed at low gas flow rates. This re-invasion rule is based on the terminal thresholds (T_t):

$$T_t = P_t + P_w, \tag{16}$$

where P_t is the terminal pressure calculated from the P_e-to-P_t ratio (α), which can be obtained from the characteristic drainage and imbibition curves of the corresponding porous medium, including capillary pressure hysteresis [49]:

$$P_t = \alpha P_e, \tag{17}$$

where α accounts for capillary hysteresis between drainage and imbibition [50]. Water re-invades the gas-occupied blocks if

$$T_{t,g} > T_{e,w}, \tag{18}$$

where the subscripts g and w stand for the gas- and water-occupied blocks, respectively. When the re-invasion of water occurs on the peripheral blocks of the gas cluster, mobilization of the gas cluster occurs. If the re-invasion of water disconnects the gas clusters, a fragmentation of the gas cluster occurs. Therefore, gas invasion can only occur at a block connected to the gas cluster containing the gas-injection port. Hence, the other gas clusters can have a re-arrangement of blocks, but no further growth. This mimics the trapping of the gas phase at this scale.

The model parameters used in this study are given in Table 1. Out of the list of model parameters, P_c per block calculated using P_d (Eq. 13) is surely an uncertain parameter to enter into $\boldsymbol{\theta}$. We assume that the other parameters are known in this study.

3.2 Gas injection experimental data

We use the experimental data from a quasi-2D gas injection experiment in an acrylic glass cell of dimensions $250mm \times 250mm \times 10mm$ filled with homogeneous, water-saturated sand of 0.7 mm average grain size from the set of experiments in [34]. The gas is injected at a rate of 0.1ml/min (Experiment 0.1-A of [34]). At this injection rate, gas migrates along with fragmenting and coalescing events on their way. The discontinuous nature of the gas flow is further confirmed by continuously measuring the pressure at the injection point during the experiment [34]. The experimental setup, data collection, and processing are described in detail in [34]. We present a brief overview of the experiment we use in this study.

The experimental data is a time series of 2D binary images (around 10,000 images) obtained using a light transmission technique [53–55]. The images are obtained at the rate of 30 frames per second for a total of 330s. Optical density (OD) [55] values are used to detect gas in

Table 1 Parameter values used in the SIP model (table taken from [24])

Parameter	Symbol	Values	Units
Density of water	ρ_w	1000	kg/m^3
Acceleration due to gravity	g	9.82	m/s^2
Average $P_t - P_e$ ratio	α	0.6	– [51]
Displacement pressure	P_d	8.66	cm of H_2O [52]
Pore-Size distribution index	λ	5.57	– [52]
Model domain size	$L_x - L_z$	250 × 250	mm^2
Block discretization	$\Delta_x - \Delta_z$	1 × 1	mm^2
Cell selection weighting factor	c	[5, 15, 100]	–

the system. Gas is considered to be present in a block above an OD threshold value of 0.02.

An ideal data set for our method would be where each atomic step (individual invasion events or re-invasion events for each block) is separately visible in time. We pick this particular data set because of its high resolution in both space and time. However, the data obtained is not free from some challenges that we need to overcome to use MFP to evaluate the BME for the SIP model.

- Firstly, non-atomic events are observed in the data set even at this high temporal resolution. This means that, from one time step to the next step, multiple atomic events (e.g. invasions, re-invasions) are found to occur so that their exact sequence is not given uniquely.
- Secondly, at some time steps, the experimental data shows re-invasion at a block not as expected by the model's deterministic re-invasion rule as specified in Eq. 18.
- Thirdly, at some time steps, invasion of gas occurs at a block that does not appear to be connected to the cluster containing the original gas injection block, violating the SIP model's assumptions. This disconnection could result from the data's optical detection limits.
- Fourthly, in some time steps, the number of gas pixels also decreases from the previous time step. This violates the mass conservation principle that the model (in the absence of a variable gas density) simplifies to a volume balance.

With the current configuration of the SIP model, using MFP would thus lead to zero-probability events because of the aforementioned observations (second to fourth) in the data set. This would lead to, within the scope of the present work, meaningless BME computations. This is not an artefact of MFP but would also occur in all other methods to compute BME. The MFP is able to map it to individual zero-probability events, while other BME computation methods would merely return an overall zero value for the entire inner integral of Eq. 5.

That is why we first test our method on synthetic data, as will be discussed in the next section. We then slowly introduce the above-mentioned irregularities in our synthetic data set, and we also introduce some workarounds (discussed later) to be able to use MFP despite the non-ideal data set and the non-ideal model assumptions. The first problem leads to an extension of MFP towards non-atomic data events, while the second problem leads to an augmented probabilistic interpretation of zero-probability events. After that, we use the longest sub-sequence of the real-experimental data with non-decreasing $n_{\text{invasions}} \geq n_{\text{re-invasions}}$ number of invaded gas blocks (7 steps between image number 239 and 246), which excludes the third and fourth problem. Within that sub-sequence, the workarounds can be implemented without computational difficulties.

3.3 Synthetic data

We begin with testing our method on the SIP model with synthetic data. By using synthetic data, we first test our method under ideal conditions. Next, we introduce irregularities in the data set in a controlled manner. To that purpose, we run the SIP model, with $c = 15$ on a particular invasion threshold field ($T_{e,syn}$) with no re-invasion events (i.e. the rule given by Eq. 18 is removed) and use the results instead of a real data set. Thus, our synthetic data set consists of a sequence of atomic events and no measurement errors. It is now guaranteed that the models can follow the data with a non-zero probability. Figure 4(a) shows this synthetic truth.

As a next step, we add non-atomicity to the synthetic data set, which is also observed in the real data set (see Section 3.2). To make non-atomic synthetic data, we regularly omit time steps, such that the SIP model would need $n_{ev} = 2, 3, 6$ iterations to get from one state to the next. This means we keep only every second, third, and sixth state from our atomic data set.

3.4 Implementation of the MFP on the SIP model

In this section, we discuss the setup and implementation of the MFP on the SIP model for both a synthetic data set and the sub-sequence from a real data set. First, we describe the common implementation setup for both types of data sets. Then, in Sections 3.4.1 and 3.4.2 we will highlight the difference in scenario setups for the corresponding data set.

We choose three cell selection weighting factors to correspond to one rather random ($c = 5$), one more deterministic ($c = 100$) and one model version in between ($c = 15$) of the SIP model. This results in three model versions. The choice of these three different cell selection weighting factors can be thought of as representative sand pack experiments with different force-dominated regimes: viscous ($c = 5$) or capillary ($c = 100$) [34]. The invasion threshold field makes up the uncertain parameters $\boldsymbol{\theta}_k$ over which we have to marginalize the inner integral (random gas-invasion decisions of SIP) through the outer integral of Eq. 5 to obtain BME. Figure 4 visualizes the randomness of the SIP model with 10 model runs for each of the model versions (c = 5 or 15 or 100). In contrast to the didactic example of Section 2.3 (Fig. 2), for SIP model we consider at each "node" of the decision tree a random decision of gas migration. These random decisions each have multiple, situation-specific possibilities for invasion and re-invasion (illustrated in Fig. 5) instead of binary decisions.

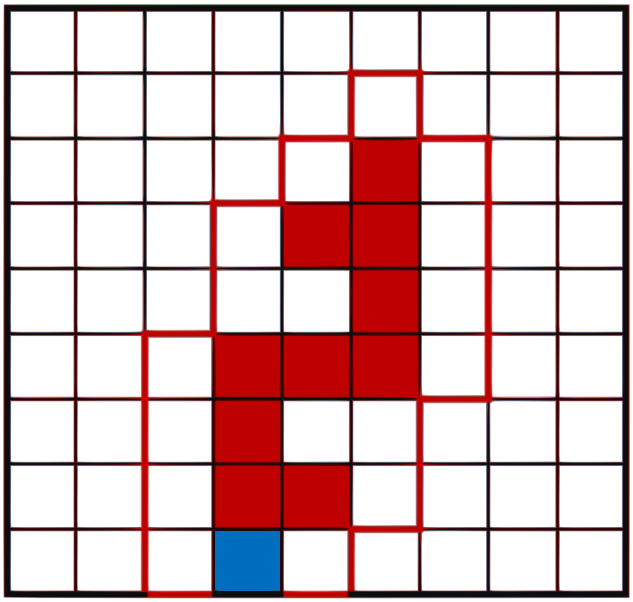

Fig. 5 Schematic to visualize the SIP model (especially their difference to the didactic example in Fig. 2); The blue block marks the injection block. The red-filled blocks mark the currently invaded blocks. In the next step, any block on the interface (red-rimmed blocks) might be invaded, and any one of the red-filled blocks or none might be re-invaded

Also, unlike the didactic example in Section 2.3, the probabilities for each forced time-step in SIP will not be a constant π_k, or $1 - \pi_k$, but they will depend on multiple factors, namely: (1) the cell selection weighting factor (c), (2) the invasion threshold (T_e) field that depends on the randomized P_c fields and, (3) the current shape of the cluster of gas-invaded blocks at the current time-step. We would like to recall, from Section 3.1, the cumulative sum $T_{e,cum}$ (Eq. 14) and its connection to the uniformly distributed random variable $\mathcal{R}$ in Eq. 15. The model chooses to invade the block with the index i (of the ascending order structure) and not any other neighbouring block if and only if Eq. 15 is fulfilled for i and not for $i - 1$, i.e.,

$$T_{e,cum}[i-1] \leq \mathcal{R}^c \times \sum_{j=1}^{j=n} T_e[j] < T_{e,cum}[i] \tag{19}$$

Rearranging the terms in the equation above gives us two bounds, and $\mathcal{R}$ must be between

$$\left(\frac{T_{e,cum}[i-1]}{\sum_{j=1}^{j=n} T_e[j]}\right)^{\frac{1}{c}} \leq \mathcal{R} < \left(\frac{T_{e,cum}[i]}{\sum_{j=1}^{j=n} T_e[j]}\right)^{\frac{1}{c}}. \tag{20}$$

The interval width between these bounds is the probability of this exact invasion at the block i. Note here that, for the block with index $i = 1$, the lower bound remains undefined by Eq. 14 and is set to zero.

When we investigate non-atomic steps in the data, we cannot simply evaluate the transition kernel of our model implied by Eq. 20 but must think about a workaround. One can view it like an excerpt of a complete probability tree diagram as shown in Fig. 2, where n_{ev} atomic events occur. The difficulty lies in not knowing the states and their ordering in between. We know the start and the end and can only guess the sequence of atomic events that happened in between. This means that any permutation of the events could be a suitable choice.

A reasonable treatment is to consider all the permutations, i.e. compute the BME over all these possible permutations. Only permutations that do not lead to a path the model can traverse by its underlying rules (see Section 3.1) and give by definition a probability of zero can be excluded. Moreover, further, to not favour any specific one of the remaining permutations, it is reasonable (and statistically

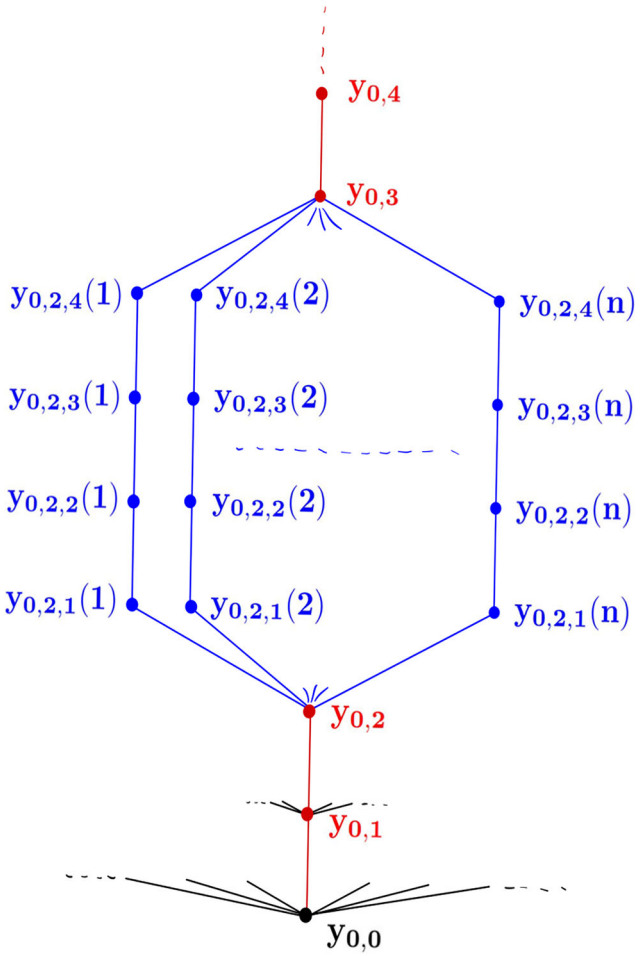

Fig. 6 Schematic for the workaround for non-atomic data steps. In the experiment path $\mathbf{y_{0,a}}$, the step from $a = 2$ to $a = 3$ consists of 5 atomic events. Each blue path corresponds to one out of $\mathbf{n}$ paths of atomic events leading to $\mathbf{y_{0,3}}$, with $\mathbf{n}$ being the number of possible permutations of the order of the atomic events. (Here, in the schematic it is $\mathbf{n} = 5! = 120$)

correct) to average their BME (see Fig. 6). We call this workaround a mini-Monte Carlo (mini-MC) approach.

If we assume that the number n_{ev} of the non-atomic events is bounded by a constant m throughout the whole experiment, we increase computational effort by a factor of $m!$, but we preserve the linear complexity of our method in t_{max} as mentioned in Section 2.3. It is reasonable to assume $m \ll t_{max}$, and thus we only employ an exhaustive search on a small scale and do not affect the overall effort significantly.

3.4.1 Synthetic scenarios

This section specifies scenario setups to treat the synthetic data set from Section 3.3. We split up our evaluations into three synthetic data scenarios as follows.

Scenario 1: In this scenario, we plug in the true invasion threshold field from Section 3.3 (i.e. the field $T_{e,syn}$ used to generate the synthetic data set) for all 3 model versions (visualization in Fig. 4(b)) and evaluate BME with our method on the atomic synthetic data and the non-atomic synthetic data (i.e, with 2, or 3, or 6 − step jumps). This scenario represents gas-injection experiment repetitions in the same sand pack without any disturbances to the setup (ideally).

Scenario 2: In this scenario, we draw an ensemble of 1000 invasion threshold fields by adding small, random noise to the true invasion threshold field ($T_{e,syn}$). Then, we plug each of these fields into the 3 model versions (visualization in Fig. 4(c)) for both the atomic and non-atomic synthetic data (with 2, or 3 − step jumps). This scenario represents gas-injection experiment repetitions in the same sand pack with smoothed-out local heterogeneities or disturbances, e.g. due to grain re-arrangement during the injection of gas.

Scenario 3: This scenario involves an ensemble of 1000 independent random invasion threshold fields, each of which is plugged into the 3 model versions (visualization in Fig. 4 (d)) for both the atomic and non-atomic synthetic data (with 2,or 3 − step jumps). This scenario represents gas-injection experiment repetitions, where the sand is repacked after each experiment.

3.4.2 Real data scenario

When using the real data sequence as mentioned in Section 3.2, we use a setup similar to Scenario 3 of Section 3.4.1 with 7000 random invasion threshold fields. The difference being that we now use the SIP model with the ability for re-invasions (recall that Section 3.4.1 uses SIP model without Eq. 18), so that they can better resemble the real data set from Section 3.2. An immediate evaluation of these models leads to BME=0 for almost all invasion threshold fields; because of its deterministic re-invasion decision (rule specified in Eq. 18), the model wants to re-invade a wrong block and is punished with complete Bayesian rejection. Theoretically, the BME value of 0 is correct, but it has no practical significance.

The focus of this study is primarily method development and not model development. Therefore, we probabilistically change the model. We assign a 90% probability to the model's decision to re-invade the block obtained from the rule specified in Eq. 18 or not-reinvade any block. The remaining probability of 10% is uniformly distributed among the other blocks of the gas cluster for the re-invasion of water. That means, any block of the current gas cluster can be re-invaded with a probability of at least $\frac{0.1}{n_{\text{gas,cluster}}}$, see Fig. 7. Note, that $n_{\text{gas,cluster}}$ only accounts for the blocks in the respective cluster. Also, we have to treat the injection cluster differently as we have one less choice since the injection block cannot be re-invaded.

We also need to adjust our workaround for the non-atomic data (Fig. 6) because we now have a re-invasion rule

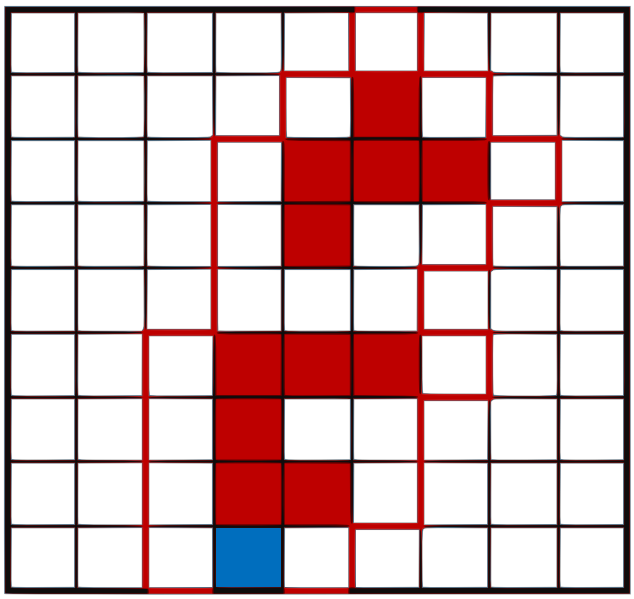

Fig. 7 Illustration of the modification to tackle the second challenge in data from Section 3.2, where water re-invasion in gas-occupied blocks occur **not** according to the model's choice (guided by Eq. 18): The blue block marks the injection block. The red blocks are gas-occupied. In this example, top gas cluster has a probability of $\frac{0.1}{n_{\text{gas,top}}} = \frac{0.1}{5}$ and the injection cluster has a probability of $\frac{0.1}{n_{\text{gas,injection}}} = \frac{0.1}{6}$, for a re-invasion of water in the respective cluster

in the models. To do that, we combine the different orderings of re-invasions with the orderings of invasions from before and leave the rest of the non-atomic modification unchanged. Note, that an atomic time-step may also have no re-invasion at all, since $n_{\text{invasions}} \geq n_{\text{re-invasions}}$. The total number of orderings is then $n_{\text{orderings}} = (n_{\text{invasions}}!)^2/(n_{\text{invasions}} - n_{\text{re-invasions}})!$. For example, a combination of non-atomic events with 5 invasions and 2 re-invasions leads to 2400 different orderings, which happens to be the maximum number for the sub-sequence mentioned in Section 3.2

3.5 MFP: list of algorithmic steps

Before we discuss the results from our case study, we summarize the general algorithmic steps of the MFP. These steps are the same for all models obeying the Markov Chain property combined with exact data (knowledge of each atomic event).

(1) List all possible events in the data, both reproducible and non-reproducible, by the model. For example, in the case of our demonstration case study, the data's non-atomic events fall under the model non-reproducible events category.
(2) State the formula for the probabilities of events being executed by the model. These could be individual, fixed values, evaluations of a probability distribution function or a combination of both. In our case study, it is stated by Eq. 20.
(3) In the original model code, code a new update rule to force the next model state, similar to a restart capability of a code.
(4) Propagate and accumulate probabilities through all time steps, i.e. a simple multiplication. At this stage, a possible code break-off criterion can also be included to identify and flag zero-probability events.

The implementation of our code is mostly non-intrusive because no re-writing of the code is necessary. However, step (3) requires good restart abilities of the model code with forced model states per time step. Also, the simplest way to achieve step (2) is to add a line to the original code that outputs the probability of the forced event.

4 Results and discussion

This section discusses the results obtained from our analysis. Table 2 contains the BME values on a negative logarithmic scale (the smaller these values are, the better the model) obtained using MFP on the SIP model for both our synthetic-data case as well as our real-data case.

4.1 Results from the synthetic data case

In both Scenario 1 and 2, the model version with $c = 15$ has the best BME values (see bold font in Synthetic Scenario 1 and 2 of Table 2). For Scenario 1, this is what we expect because this model version and threshold field were used

Table 2 Table containing the BME values obtained in the three synthetic scenarios and the real scenario on a negative logarithm scale, the ensemble sizes n_{MC}, number of atomic events n_{ev} occurring within a non-atomic step and, computed Bayes factors $\text{BF}_{\frac{k2}{k1}}$

Scenario		n_{MC}	n_{ev}	$c = 5$	$c = 15$	$c = 100$	$\text{BF}_{\frac{15}{5}}$	$\text{BF}_{\frac{5}{100}}$	$\text{BF}_{\frac{15}{100}}$
Synthetic	1	1	1	3034.6	**2672.3**	3063.3	2.6e157	3.0e12	6.1e169
			2	3099.9	**2741.0**	3124.6	7.5e155	5.3e10	4.0e166
			3	3169.8	**2817.5**	3202.3	9.7e152	1.3e14	1.3e167
			6	3301.9	**2961.8**	3345.5	4.8e147	9.1e18	4.4e166
	2	1000	1	3201.2	**2931.1**	3474.1	2.1e117	3.4e118	7.1e235
			2	3246.7	**2972.1**	3496.0	1.7e119	2.0e108	3.4e227
			3	3292.6	**3013.2**	3519.0	2.3e121	2.1e98	5.0e219
	3	1000	1	**5942.7**	6647.7	8461.9	6.4e − 307	1.2e1094	8.0e787
			2	**5942.1**	6644.5	8453.3	9.0e − 306	3.9e1090	3.5e785
			3	**5939.6**	6638.3	8439.8	3.6e − 304	7.0e1085	2.5e782
Real		7000	3 − 5	**269.22**	294.29	336.40	1.3e − 11	1.5e29	1.9e19

Note, here the model versions ($k1, k2$), are denoted by their respective c values. The best performing model is highlighted with bold font − ln **BME** value

to generate the synthetic data. For Scenario 2, the threshold fields were close to the synthetic data setup; therefore, the correct model version still had the best BME value. Also, according to expectations, all the model versions had significantly worse BME values for Scenario 3, where entirely random entry threshold fields were used. However, the ranking also changed, and the more random model version ($c = 5$) emerged as the best model in Scenario 3.

4.1.1 Why does the model ranking change for Scenario 3?

Let us first look at the two extreme model versions to understand why the ranking changes. The model version with $c = 100$ is almost deterministic in its choice of a gas pathway, which is different for each invasion threshold field. This is why this model version can get good BME values (small $-\ln BME$) if and only if the invasion threshold field closely matches the true field ($T_{e,syn}$), which is highly unlikely when we use entirely random invasion threshold fields. If this is put colloquially, the few good predictions of the $c = 100$ model version do not make up for the many bad ones. The more random ($c = 5$) model version is not as deterministic in its choice of the gas pathway as $c = 100$ is, and so, it is largely unimpaired by the choice of the invasion threshold field. This is why the random model version ($c = 5$) achieves mediocre values for any invasion threshold field. Thus, in the scenario where the invasion threshold field is highly uncertain, it has an advantage that helps it emerge as the best model version in Scenario 3. The model version with $c = 15$ is not identified as the best model when we increase the uncertainty in the threshold field. This indicates that the entry threshold field is a highly sensitive and important parameter for the SIP model to function correctly.

4.1.2 Effect of non-atomic synthetic data

The introduction of non-atomicity in the synthetic data does **not** change the ranking of the models in any scenario (see, $-$ lnBME values for n_{ev} values other than 1 for Synthetic Scenarios (1, 2, and 3) in Table 2) but makes it slightly less decisive in comparison to $n_{ev} = 1$ for all the synthetic scenarios of Table 2. This coincides with the synthetic data set becoming, in a sense, weaker or less informative if parts of it are unknown in ordering. This is visible in Table 2 as, despite the general rise of $-$ ln BME values with increasing n_{ev}, their differences become slightly smaller. Looking at the Bayes factors between the competing models makes it easier to see this effect: they generally decrease with increasing n_{ev}. There are only a few exceptions to this observation. For example, the Bayes Factors $BF_{\frac{15}{5}}$ between the models in Scenario 2 increases with the increased non-atomicity in the synthetic data. However, looking at the orders of magnitude of the values in comparison, it can be safely concluded that this does not affect or change the level of decisiveness.

4.2 Results from real data scenario

Initially, we evaluate the model versions on the complete real data set. This helps us gather information on the magnitude of the effect of the challenges in the real data for implementation of MFP, as discussed in Section 3.2. We find that non-atomic events with a very high number of events n_{ev} are pretty common in the data set, which leads to very high computation time for the mini-MC workaround explained in Section 3.4, thus making a BME evaluation infeasible even with MFP.

The events of wrong block re-invasion (the second problem in data discussed in Section 3.2) in the data set are plenty, but we are able to tackle them with our workaround mentioned in Section 3.4.2. In the later time-steps of the data, block invasion in non-gas injection clusters (Third problem in data discussed in Section 3.2) or events with decreasing numbers of invaded gas blocks (Fourth problem in data discussed in Section 3.2) are predominant. However, we have no fix to this problem in the real data set.

Thus, we decide to look for a sub-sequence of time steps in the data that aligns with the model's assumptions and has a reasonably small number of non-atomic events (for reasonable computation time of mini-MC runs, see Fig. 6). The resulting sequence of time steps in the data is the one mentioned in Section 3.2. For that sequence, $-$ lnBME is between 269 and 360 for the probability of correctly predicting seven steps, which is a small probability already. This is not a fault of our method MFP. We hope that the models to which our readers may decide to apply MFP match their corresponding data set better than in our case.

Regarding the ranking of the model versions, we see a similar pattern as in the synthetic Scenario 3. The best model version is the one with $c = 5$, followed by $c = 15$ and then $c = 100$ (see Row: Real from Table 2). The uncertainty in the invasion threshold fields is handled better by a random (c=5) model than by the more deterministic models. Therefore, more information about the invasion threshold fields is necessary for these models to accurately predict the gas path under the experimental data's conditions and scale.

5 Conclusions and outlook

In conclusion, our method MFP makes it possible to calculate BME for Markov-Chain type models and discrete atomic data in previously impossible cases. The method works well, is non-intrusive to the model and has a linear

computational cost. In our case study, the method was demonstrated only on a relatively small sequence of real data. This is because the large distance between the model outputs and the real data leads to many zero-probability events. So, for more conclusive results, better models or more-informative data are required, i.e. data with no or few non-atomic events and an improved SIP model.

When we use MFP to evaluate the BME for imperfect models or data or both, resulting in practically futile BME values (BME = 0), we can adapt our approach and use MFP to detect events leading to such values in the model and the data by flagging them. This exercise helps determine the structural errors in the model, or mismatch between the model concepts and the observations.

From our implementation of MFP on the SIP model and gas-injection experimental data used in this study, we can conclude that both the model and the experimental data have a scope for improvement. The rules in the SIP model could be updated by looking at specific types of events, e.g. the ones that get the model rejected or result in poor performance (like the deterministic re-invasion events in the current version). The experimental data technique processing could be updated to have more discrete and atomic data steps. However, experimental data or model improvement is beyond the scope of the present research work.

Our method, in its current stage of development, requires that the data be noise-free. Further research is needed to apply MFP to noisy data (e.g. with statistical assumptions on the distribution of black/white detection errors). A straightforward idea would be to perturb the available data with several realizations of randomly generated noise and then handle each realization with our method. However, this multiplies computational costs by a substantial factor to host these repetitions.

Our method enables a fully Bayesian assessment of the SIP model for the first time. Besides the SIP model and gas-injection experimental data, this method can be applied to gas (fluid) migration in fractured-porous media under the conditions of Markov-style model formulation and complete observations (e.g. in thin slices or with high-resolution 3D micro-tomography). It can also be applied to systems involving experiments and models at the microscopic scale, where individual pores are resolved by appropriate monitoring techniques (e.g., [56]).

Apart from multiphase flow in porous media problems, this method can be used in applications such as counting processes (e.g. as in traffic), discrete computerized systems (e.g. network traffic), probabilistic Markov-style model-based river water quality monitoring [57], tracer experiments / Lagrangian movement (e.g. fluorescent microparticles to monitor turbulent flow [58]), stochastic models for discrete, dynamic systems and complete observation (e.g. chemical reaction modelling), statistics-based data-driven soil-plant-atmosphere modelling [59] and micro-seismic modelling [60] to name a few.

Appendix A: Monte Carlo simulations grow exponentially in t_{max}

We will prove here that the required size of MC simulations grows exponentially in t_{max}. Based on the tree structure in the didactic example from Section 2.3, we can see that there are $N = 2^{t_{max}}$ equiprobable branches. Therefore, the probability of the correct branch (the one with non-zero likelihood) is exactly:

$$P_{true} = \frac{1}{N} = \frac{1}{2^{t_{max}}} = 2^{-t_{max}} = BME, \tag{21}$$

which apparently is the situation-specific definition of BME. When approximating BME via MC sampling with $i = 1 \ldots n$ independent random realizations, then each realization i has a constant and independent probability of finding or not finding the correct branch. This situation is described exactly by the Binomial distribution. The Binomial distribution is a discrete probability distribution of the number of success events k out of n independent trials:

$$P(X = k) = \binom{n}{k} p^k \, (1-p)^{(n-k)} \tag{22}$$

where p is the probability of success and $1 - p$ is the probability of not obtaining success. Here, we define n as number of MC trials, X is the (random under repetition of the entire MC simulation) number of times the MC finds the correct branch (i.e., the one with $Likelihood = 1$). For a given execution of MC with given n, we will find k times the correct branch and $n - k$ times any branch with $Likelihood = 0$.

Also, in the context of our example, the probability parameter p of the Binomial distribution is equal to $BME = P_{true}$. From the MC results, one would estimate:

$$\widehat{BME}_{MC} = \frac{k}{n} = \hat{p} \approx BME$$

Just for reassurance, the asymptotic MC result for BME at $n \to \infty$ converges to the exact solution:

$$\mathrm{E}\left[\frac{X}{n}\right] = \frac{1}{n} E\left[X\right] = p = BME.$$

But can we estimate the MC error in this approximation, e.g., expressed as the coefficient of variation (CV). For the Binomial distribution, we know that:

$$\text{Variance of } X\text{: } \mathrm{Var}[X] = n \cdot p \cdot (1-p)$$
$$\text{Mean of } X\text{: } \mathrm{E}[X] = n \cdot p$$

As the conversion from k to the estimate of p is simply a division by n, we apply the rules of linearized uncertainty quantification to see that:

$$\text{Variance of } \frac{X}{n}\text{: } \mathrm{Var}\left[\frac{X}{n}\right] = \frac{p\cdot(1-p)}{n}$$
$$\text{Mean of } \frac{X}{n}\text{: } \mathrm{E}\left[\frac{X}{n}\right] = p$$

Using this in the definition of the coefficient of variation:

$$\text{CV of } \frac{X}{n} = CV\left[\widehat{BME}\right] = \frac{\sqrt{\mathrm{Var}\left[\frac{X}{n}\right]}}{\mathrm{E}\left[\frac{X}{n}\right]} = \frac{\sqrt{1-p}}{\sqrt{n}\cdot\sqrt{p}}$$

For small values of p as in the given example with $p = 2^{-t_{max}}$, we can replace $1-p \approx 1$, and hence:

$$\text{CV of } \frac{X}{n} \text{ for small } p\text{: } CV = \frac{1}{\sqrt{n}\cdot\sqrt{p}} = \frac{1}{\sqrt{n}\cdot\sqrt{2^{-t_{max}}}} = \frac{2^{\frac{t_{max}}{2}}}{\sqrt{n}}.$$

Thus, the number of MC runs required for a desired accuracy (expressed as a desired value of the CV) is:

$$n_{required} = \frac{2^{t_{max}}}{CV^2_{desired}}.$$

This shows that the number n of required MC samples increases, for a given precision requirement, exponentially in t_{max}. The base 2 of the exponent originates from the tree structure, where each node expands into two further branches. In real applications, where the evolution of the model over time has more than two possibilities, the base will simply increase, so the exponential growth will be even stronger.

Appendix B: Tackling numerical instabilities in computation of *BME*

Here, we provide details on the approach of handling numerical instabilities in the BME computation when using MC integration (Eq. 2) for the uncertain parameters in Eq. 5.

To avoid very small likelihoods (BME values from the perspective of random events $\boldsymbol{\omega}$) turning into numerical zeros, we divide each sample likelihood by the maximum likelihood encountered in the whole ensemble, $\max\{p(\mathbf{y}_0 \mid \boldsymbol{\theta}_k, M_k)\}$, yielding values between 0 and 1. Then, Eq. 5 rewrites as:

$$I_k = \max\{p(\mathbf{y}_0 \mid \boldsymbol{\theta}_k, M_k)\} \int \frac{p(\mathbf{y}_0|\boldsymbol{\theta}_k, M_k)}{\max\{p(\mathbf{y}_0|\boldsymbol{\theta}_k, M_k)\}} \cdot p\left(\boldsymbol{\theta}_k \mid M_k\right) d\boldsymbol{\theta}_k.$$

Taking the logarithm and applying the MC approximation of the integral yields:

$$\ln I_k = \ln \max\{p(\mathbf{y}_0 \mid \boldsymbol{\theta}_k, M_k)\} - \ln N + \ln \sum_{r=1}^{N} \frac{p(\mathbf{y}_0|\boldsymbol{\theta}_{k,r}, M_k)}{\max\{p(\mathbf{y}_0|\boldsymbol{\theta}_k, M_k)\}}.$$

Acknowledgements The authors would like to thank the German Research Foundation (DFG) for financial support of this project within the Research Training Group GRK1829 "Integrated Hydrosystem Modelling" and the Cluster of Excellence EXC 2075 "Data-integrated Simulation Science (SimTech)" at the University of Stuttgart under Germany's Excellence Strategy - EXC 2075 - 39074001. The authors would also like to thank Assistant Professor Dr. Cole Van De Ven, Carleton University, Canada, for his assistance with the provisioning, handling and processing of the experimental data used in this study.

Author Contributions 'Not applicable'

Funding Open Access funding enabled and organized by Projekt DEAL. The authors would like to thank the German Research Foundation (DFG) for financial support of this project within the Research Training Group GRK1829 "Integrated Hydrosystem Modelling" and the Cluster of Excellence EXC 2075 "Data-integrated Simulation Science (SimTech)" at the University of Stuttgart under Germany's Excellence Strategy - EXC 2075 - 39074001.

Data Availability The experimental data set used in this study is made available by [61] at Scholars Portal Dataverse: https://doi.org/10.5683/SP2/RQKOCN.

Code Availability The modelling data and codes used in this study are available as a data set [62] in the DaRUS dataverse for Stochastic Simulation and Safety Research for Hydrosystems (LS3):https://doi.org/10.18419/darus-2815.

Declarations

Ethics approval and consent to participate 'Not applicable'

Consent for Publication 'Not applicable'

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Schöniger, A., Wöhling, T., Samaniego, L., Nowak, W.: Model selection on solid ground: rigorous comparison of nine ways to evaluate Bayesian model evidence. Water Resour. Res. **50**, 9484–9513 (2014). https://doi.org/10.1002/2014WR016062
2. Kashyap, R.L.: Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-4**, 99–104 (1982). https://doi.org/10.1109/TPAMI.1982.4767213
3. Gideon, S.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
4. Hirotugu, A.: Information theory and an extension of the maximum likelihood principle. Paper presented at the Second International Symposium on Information Theory, Czaki. Akademiai Kiado, Budapest (1973)
5. Poeter, E., Anderson, D.: Multimodel ranking and inference in ground water modeling. Ground Water. **43**, 597–605 (2005). https://doi.org/10.1111/j.1745-6584.2005.0061.x
6. Ye, M., Meyer, P.D., Neuman, S.P.: On model selection criteria in multimodel analysis. Water Resour. Res. 44. https://doi.org/10.1029/2008WR006803 (2008)
7. Ye, M., Pohlmann, K.F., Chapman, J.B., Pohll, G.M., Reeves, D.M.: A model-averaging method for assessing groundwater conceptual model uncertainty. Ground Water. **48**, 716–728 (2010). https://doi.org/10.1111/j.1745-6584.2009.00633.x
8. Ye, M., Lu, D., Neuman, S.P., Meyer, P.D.: Comment on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window" by Frank T.-C. Tsai and Xiaobao Li. Water Resour. Res., 46. https://doi.org/10.1029/2009WR008501 (2010)
9. Tsai, F.TC., Li, X.: Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window. Water Resour. Res., 44. https://doi.org/10.1029/2007WR006576 (2008)
10. Tsai, F.TC., Li, X.: Reply to comment by Ming Ye others. on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window". Water Resour. Res., 46. https://doi.org/10.1029/2009WR008591 (2010)
11. Singh, A., Mishra, S., Ruskauff, G.: Model averaging techniques for quantifying conceptual model uncertainty. Ground Water. **48**, 701–715 (2010). https://doi.org/10.1111/j.1745-6584.2009.00642.x
12. Morales-Casique, E., Neuman, S.P., Vesselinov, V.V.: Maximum likelihood Bayesian averaging of airflow models in unsaturated fractured tuff using Occam and variance windows. Stoch. Environ. Res. Risk Assess. **24**, 863–880 (2010). https://doi.org/10.1007/s00477-010-0383-2
13. Foglia, L., Mehl, S.W., Hill, M.C., Burlando, P.: Evaluating model structure adequacy: the case of the Maggia Valley groundwater system, southern Switzerland. Water Resour. Res. **49**, 260–282 (2013). https://doi.org/10.1029/2011WR011779
14. Kloek, T., van Dijk, H.K.: Bayesian estimates of equation system parameters. An application of integration by Monte Carlo. Econometrica **46**, 1–19 (1978). https://doi.org/10.2307/1913641
15. Zellner, A., Rossi, P.E.: Bayesian analysis of dichotomous quantal response models. J. Econom. **25**, 365–393 (1984). https://doi.org/10.1016/0304-4076(84)90007-1
16. Snyder, C., Bengtsson, T., Bickel, P., Anderson, J.: Obstacles to high-dimensional particle filtering. Mon. Weather Rev. **136**, 4629–4640 (2008). https://doi.org/10.1175/2008MWR2529.1
17. Bengtsson, T., Bickel, P., Li, B.: Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In: Nolan, D., Speed, T. (eds.) Probability and Statistics: essays in Honor of David A. Freedman, vol. 2, pp. 316-334. Institute of Mathematical Statistics, Beachwood (2008)
18. Zhang, Y., Liu, Y., Pau, G., Oladyshkin, S., Finsterle, S.: Evaluation of multiple reduced-order models to enhance confidence in global sensitivity analyses. Int. J. Greenhouse Gas Control. **49**, 217–226 (2016). https://doi.org/10.1016/j.ijggc.2016.03.003
19. Xiu, D., Karniadakis, G.E.: The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002). https://doi.org/10.1137/S1064827501387826
20. Willcox, K., Peraire, J.: Balanced model reduction via the proper orthogonal decomposition. AIAA J. **40**(11), 2323–2330 (2002). https://doi.org/10.2514/2.1570
21. Kumar, R., Tolson, B.A., Burn, D.H.: Review of surrogate modeling in water resources. Water Resour. Res. 48(7). https://doi.org/10.1029/2011WR011527 (2012)
22. Beaumont, M.A.: Approximate Bayesian Computation in Evolution and Ecology. Annual Review of Ecology. Evolution, and Systematics. **41**(1), 379–406 (2010). https://doi.org/10.1146/annurev-ecolsys-102209-144621
23. Mumford, K.G., Hegele, P.R., Vandenberg, G.P.: Comparison of two-dimensional and three-dimensional macroscopic invasion percolation simulations with laboratory experiments of gas bubble flow in homogeneous sands. Vadose Zone J. **14**, 1–13 (2015). https://doi.org/10.2136/vzj2015.02.0028
24. Banerjee, I., Guthke, A., Van De Ven, C.JC., Mumford, K.G., Nowak, W.: Overcoming the model-data-fit problem in porous media: a quantitative method to compare invasion-percolation models to high-resolution data. Water Resour. Res. **57**, e2021WR029986 (2021). https://doi.org/10.1029/2021WR029986
25. Kass, R.E., Raftery, AE.: Bayes factors. J. Amer. Stat. Assoc. **90**, 773–795 (1995). https://doi.org/10.1080/01621459.1995.10476572
26. Hammersley, J.M.: Monte Carlo Methods for solving multivariable problems. Ann. New York Acad. Sci. **86**, 844–874 (1960). https://doi.org/10.1111/j.1749-6632.1960.tb42846.x
27. Jeffreys, H. Theory of Probability, 3rd edn. Clarendon Press, Oxford (1961)
28. Kolmogorov, A.: Foundations of the Theory of Probability. Morrison N, editor. Chelsea Publishing Company, New York (1950)
29. Gustafsson, F.: Particle filter theory and practice with positioning applications. IEEE Aerosp. Electron. Syst. Mag. **25**(7), 53–82 (2010). https://doi.org/10.1109/MAES.2010.5546308
30. Ramgraber, M., Albert, C., Schirmer, M.: Data assimilation and online parameter optimization in groundwater modeling using nested particle filters. Water Resour. Res. **55**(11), 9724–9747 (2019). https://doi.org/10.1029/2018WR024408
31. Nowak, W.: Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator. Water Resour. Res. 45(4). https://doi.org/10.1029/2008WR007328 (2009)
32. Schöniger, A., Nowak, W., Hendricks Franssen, H.J.: Parameter estimation by ensemble Kalman filters with transformed data: approach and application to hydraulic tomography. Water Resour. Res. 48(4). https://doi.org/10.1029/2011WR010462 (2012)
33. Ewing, R.P., Berkowitz, B.: A generalized growth model for simulating initial migration of dense non-aqueous phase liquids. Water Resour. Res. **34**, 611–622 (1998). https://doi.org/10.1029/97WR03754
34. Van De Ven, C.JC., Mumford, K.G.: Characterization of gas injection flow patterns subject to gravity and viscous forces. Vadose Zone J. **18**, 1–11 (2019). https://doi.org/10.2136/vzj2019.02.0014
35. Broadbent, S.R., Hammersley, J.M.: Percolation processes. Mathematical Proceedings of the Cambridge Philosophical Society. https://doi.org/10.1017/s0305004100032680 (1957)

36. Wilkinson, D., Willemsen JF.: Invasion percolation: a new form of percolation theory. J. Phys. A: Math. Gen. **16**, 3365–3376 (1983). https://doi.org/10.1088/0305-4470/16/14/028
37. Wilkinson, D.: Percolation model of immiscible displacement in the presence of buoyancy forces. Phys. Rev. A. **30**, 520–531 (1984). https://doi.org/10.1103/PhysRevA.30.520
38. Birovljev, A., Furuberg, L., Feder, J., Jssang, T., Mly, K.J., Aharony, A.: Gravity invasion percolation in two dimensions: experiment and simulation. Phys. Rev. Lett. **67**, 584–587 (1991). https://doi.org/10.1103/PhysRevLett.67.584
39. Birovljev, A., Wagner, G., Meakin, P., Feder, J., Jøssang, T.: Migration and fragmentation of invasion percolation clusters in two-dimensional porous media. Phys. Rev. E. **51**, 5911–5915 (1995). https://doi.org/10.1103/PhysRevE.51.5911
40. Kueper, B.H., McWhorter, D.B.: The use of macroscopic percolation theory to construct large-scale capillary pressure curves. Water Resour. Res. **28**, 2425–2436 (1992). https://doi.org/10.1029/92WR01176
41. Wagner, G., Meakin, P., Feder, J., Jøssang, T.: Buoyancy-driven invasion percolation with migration and fragmentation. Physica A: Stat. Mech. Applic. **245**, 217–230 (1997). https://doi.org/10.1016/S0378-4371(97)00324-5
42. Ewing, R.P., Berkowitz, B.: Stochastic pore-scale growth models of DNAPL migration in porous media. Adv. Water Resour. **24**, 309–323 (2001). https://doi.org/10.1016/S0309-1708(00)00059-2
43. Glass, R.J., Conrad, S.H., Yarrington, L.: Gravity-destabilized nonwetting phase invasion in macroheterogeneous porous media: near-pore-scale macro modified invasion percolation simulation of experiments. Water Resour. Res. **37**, 1197–1207 (2001). https://doi.org/10.1029/2000WR900294
44. Mumford, K.G., Smith, J.E., Dickson, S.E.: The effect of spontaneous gas expansion and mobilization on the aqueous-phase concentrations above a dense non-aqueous phase liquid pool. Adv. Water Resour. **33**, 504–513 (2010). https://doi.org/10.1016/j.advwatres.2010.02.002
45. Trevisan, L., Illangasekare, T.H., Meckel, T.A.: Modelling plume behavior through a heterogeneous sand pack using a commercial invasion percolation model. Geomech. Geophys. Geo-Energy Geo-Resour. **3**, 327–337 (2017). https://doi.org/10.1007/s40948-017-0055-5
46. Molnar, I.L., Mumford, K.G., Krol MM.: Electro-thermal subsurface gas generation and transport: model validation and implications. Water Resour. Res. **55**, 4630–4647 (2019). https://doi.org/10.1029/2018WR024095
47. Brooks, R., Corey, A.: Hydraulic properties of porous media. Water Resour. Res., 4 (1964)
48. Zhao, W., Ioannidis, MA.: Gas exsolution and flow during supersaturated water injection in porous media: I. Pore network modeling. Adv. Water Resour. **34**, 2–14 (2011). https://doi.org/10.1016/j.advwatres.2010.09.010
49. Gerhard, J.I., Kueper, B.H.: Capillary pressure characteristics necessary for simulating DNAPL infiltration, redistribution, and immobilization in saturated porous media. Water Resources Research., 39. https://doi.org/10.1029/2002WR001270 (2003)
50. Ioannidis, M.A., Chatzis, I., Dullien, F.A.L.: Macroscopic percolation model of immiscible displacement: effects of buoyancy and spatial structure. Water Resour. Res. **32**(11), 3297–3310 (1996). https://doi.org/10.1029/95WR02216
51. Mumford, K.G., Dickson, S.E., Smith, J.E.: Slow gas expansion in saturated natural porous media by gas injection and partitioning with non-aqueous phase liquids. Adv. Water Resour. **32**, 29–40 (2009). https://doi.org/10.1016/j.advwatres.2008.09.006
52. Schroth, M.H., Istok, J.D., Ahearn, S.J., Selker, J.S.: Characterization of Miller-similar silica sands for laboratory hydrologic studies. Soil Sci. Soc. Amer. J. **60**, 1331–1339 (1996). https://doi.org/10.2136/sssaj1996.03615995006000050007x
53. Tidwell, V.C., Glass, R.J.: X ray and visible light transmission for laboratory measurement of two-dimensional saturation fields in thin-slab systems. Water Resour. Res. **30**(11), 2873–2882 (1994). https://doi.org/10.1029/94WR00953
54. Niemet, M.R., Selker, J.S.: A new method for quantification of liquid saturation in 2D translucent porous media systems using light transmission. Adv. Water Resour. **24**(6), 651–666 (2001). https://doi.org/10.1016/S0309-1708(00)00045-2
55. Kechavarzi, C., Soga, K., Wiart, P.: Multispectral image analysis method to determine dynamic fluid saturation distribution in two-dimensional three-fluid phase flow laboratory experiments. Journal of Contaminant Hydrology. https://doi.org/10.1016/S0169-7722(00)00133-9 (2000)
56. Gao, H., Tatomir, A.B., Karadimitriou, N.K., Steeb, H., Sauter, M.: Effects of surface roughness on the kinetic interface-sensitive tracer transport during drainage processes. Adv. Water Resour. **104044**, 157 (2021). https://doi.org/10.1016/j.advwatres.2021.104044
57. Gonzalez-Nicolas, A., Schwientek, M., Sinsbeck, M., Nowak, W.: Characterization of export regimes in concentration–discharge plots via an advanced time-series model and event-based sampling strategies. Water, 13. https://doi.org/10.3390/w13131723 (2021)
58. Adrian, R.J.: Twenty years of particle image velocimetry. Exper. Fluids. **39**, 159–169 (2005). https://doi.org/10.1007/s00348-005-0991-7
59. Gong, W., Gupta, H.V., Yang, D., Sricharan, K., Hero, AO. III.: Estimating epistemic and aleatory uncertainties during hydrologic modeling: an information theoretic approach. Water Resour. Res. **49**(4), 2253–2273 (2013). https://doi.org/10.1002/wrcr.20161
60. Shapiro, S.A.: Microseismicity - a Tool for Reservoir Characterization. EAGE Publications bv, Amsterdam (2008)
61. Van De Ven, C., Mumford, K., Banerjee, I.: Replication Data for: Overcoming the model-data-fit problem in porous media: a quantitative method to compare invasion-percolation models to high-resolution data. Borealis. Available from: https://doi.org/10.5683/SP2/RQKOCN
62. Banerjee, I., Walter, P.: Replication data for: The method of forced probabilities: a computation trick for Bayesian model evidence. DaRUS. Available from: https://doi.org/10.18419/darus-2815

Affiliations

Ishani Banerjee[1] · Peter Walter[1] · Anneli Guthke[2] · Kevin G. Mumford[3] · Wolfgang Nowak[1]

Peter Walter
peter-jan.walter@t-online.de

Anneli Guthke
anneli.guthke@simtech.uni-stuttgart.de

Kevin G. Mumford
kevin.mumford@queensu.ca

Wolfgang Nowak
wolfgang.nowak@iws.uni-stuttgart.de

1 Institute for Modelling Hydraulic and Environmental Systems (IWS)/LS3, University of Stuttgart, Pfaffenwaldring 5a, Stuttgart, 70569, Baden Württemberg, Germany

2 Stuttgart Center for Simulation Science, Cluster of Excellence EXC 2075, University of Stuttgart, Pfaffenwaldring, Stuttgart, 70569, Baden Württemberg, Germany

3 Department of Civil Engineering, Queen's University, Ellis Hall, Kingston, K7L 3N6, Ontario, Canada