Computational Geosciences (2021) 25:1899–1917
https://doi.org/10.1007/s10596-021-10076-9

ORIGINAL PAPER

Surrogate-based Bayesian comparison of computationally expensive models: application to microbially induced calcite precipitation

Stefania Scheurer[1] · Aline Schäfer Rodrigues Silva[1] · Farid Mohammadi[2] · Johannes Hommel[2] · Sergey Oladyshkin[1] · Bernd Flemisch[2] · Wolfgang Nowak[1]

Received: 26 November 2020 / Accepted: 7 July 2021 / Published online: 5 August 2021

Abstract

Geochemical processes in subsurface reservoirs affected by microbial activity change the material properties of porous media. This is a complex biogeochemical process in subsurface reservoirs that currently contains strong conceptual uncertainty. This means, several modeling approaches describing the biogeochemical process are plausible and modelers face the uncertainty of choosing the most appropriate one. The considered models differ in the underlying hypotheses about the process structure. Once observation data become available, a rigorous Bayesian model selection accompanied by a Bayesian model justifiability analysis could be employed to choose the most appropriate model, i.e. the one that describes the underlying physical processes best in the light of the available data. However, biogeochemical modeling is computationally very demanding because it conceptualizes different phases, biomass dynamics, geochemistry, precipitation and dissolution in porous media. Therefore, the Bayesian framework cannot be based directly on the full computational models as this would require too many expensive model evaluations. To circumvent this problem, we suggest to perform both Bayesian model selection and justifiability analysis after constructing surrogates for the competing biogeochemical models. Here, we will use the arbitrary polynomial chaos expansion. Considering that surrogate representations are only approximations of the analyzed original models, we account for the approximation error in the Bayesian analysis by introducing novel correction factors for the resulting model weights. Thereby, we extend the Bayesian model justifiability analysis and assess model similarities for computationally expensive models. We demonstrate the method on a representative scenario for microbially induced calcite precipitation in a porous medium. Our extension of the justifiability analysis provides a suitable approach for the comparison of computationally demanding models and gives an insight on the necessary amount of data for a reliable model performance.

Keywords Microbially induced calcite precipitation · Bayesian model selection · Bayesian model justifiability analysis · Arbitrary polynomial chaos expansion · Surrogate-based model selection and comparison · Surrogate-based Bayesian model justifiability analysis

1 Introduction

1.1 Biogeochemical processes in subsurface porous media

Biogeochemical processes in porous media are geochemical processes affected by the activity of microbes [37]. They profoundly impact ecosystems as they occur ubiquitously in the subsurface. This makes them interesting for applications in engineering. Some examples of biogeochemical processes that engineers tried to manipulate are: enhanced recovery of resources as in microbially enhanced oil recovery (e.g. [4, 29, 39]), blocking of preferential flow paths by the accumulation of biomass or minerals precipitated as a result of the microbial metabolism (e.g. [8, 73]), bioremediation of aquifers or soils by microbial decomposition of organic pollutants (e.g. [20, 40, 45]) or in situ sequestration of inorganic contaminants (metals, radionuclides) by biotically managed precipitation [19].

✉ Sergey Oladyshkin
sergey.oladyshkin@iws.uni-stuttgart.de

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

However, it is challenging to describe these biogeochemical processes in full detail, because many subprocesses interact in a complex manner [70]. Accordingly, it is not easy to control them as desired. A good understanding of these processes is necessary when aiming to control them in order to predict or even regulate the outcome. Thus, modeling is a crucial tool to predict the response of systems under certain conditions [30]. Corresponding models are an essential tool in investigating the coupled transport of fluids and reactive substances through porous media and the resulting chemical reactions in the pores [38, 71, 86].

Several transport models dealing with the biogeochemical process of microbially induced calcite precipitation (MICP) have been discussed in works by e.g. [5, 15, 25, 26, 46, 83]. This induced calcite precipitation provides a practical technical application. By accumulating the precipitated calcite, the porosity and permeability of a porous medium can be reduced (e.g. [13, 14, 42, 58, 72]). Additionally, MICP can be used to reduce erosion or increase soil stability (e.g. [17, 56, 80, 87]). MICP has been proven to reduce permeability and enhance mechanical strength even at large, field-relevant scales (e.g. [33, 41, 46, 56, 59]). There are several reviews about the understanding of bio-improved soils (e.g. [44, 74, 76]).

Biogeochemical models are useful, for example, to design, monitor, and evaluate such applications, e.g. to mitigate leakages from a geological gas reservoir into above aquifers in advance (e.g. [12, 13, 35, 41, 46]). Our limited knowledge about the interaction of the processes that govern biogeochemical systems leads to several modeling approaches that differ, e.g., in their level of detail. The uncertainty of choosing between these modeling alternatives is considered here as conceptual uncertainty.

1.2 Conceptual uncertainty

When modeling an environmental process, we have to make assumptions and simplifications because, usually, the real process is too complex to be represented in full detail. Consequently, one has to deal with various types of uncertainty. Besides input and parameter uncertainty, conceptual uncertainty (uncertainty of model choice) has to be taken into account. If we chose a single model and did not consider possible alternatives, we might strongly underestimate the overall prediction uncertainty because the space of potential models is not sufficiently covered [16, 61, 63].

Many studies have identified conceptual uncertainty as a key source of uncertainty in modeling (e.g [10, 16, 18, 22, 48, 61–64, 69, 75]). These studies suggest to treat modeling concepts with different levels of detail and different assumptions as competing hypotheses. By using statistical techniques such as Bayesian model selection (BMS), we can evaluate which model is the most appropriate representation of the system [60, 79].

However, two challenges persist. First, it is important to note that there is no existing method which allows to quantify conceptual uncertainty on an absolute level [24, 47]. Second, biogeochemical modeling, discussed briefly in Section 1.1, is computationally very demanding since it conceptualizes different processes in subsurface porous media. Thus, a direct application of the rigorous probabilistic machinery is not feasible due to a necessity of a high number of model evaluations. In this study, we address the second challenge.

1.3 Surrogate representation of the underlying physical models

In order to assure feasibility of the probabilistic BMS framework, we will construct computationally cheaper surrogate models for each version of the biogeochemical model. The purpose of a surrogate model is to replicate the behavior of the underlying physical model from a limited set of runs. For constructing a surrogate the original model should be evaluated by using those sets of modeling parameters out of various possibilities that cover the parametric space as well as possible. Considering very high computational cost of biogeochemical models, whereby one model evaluation requires days, we need to select an approach that will capture the main features of the underlying physical models after a very small number of model evaluations. Following a recent benchmark comparison study by [34], we construct the surrogate model using the arbitrary polynomial chaos expansion technique (aPC) introduced in [52], which is suitable for our purpose.

In short, the data-driven aPC approach can be seen as a machine learning tool that approximates the model output by its dependence on model parameters via multivariate polynomials. The data-driven feature of aPC offers complete flexibility in the choice and representation of probability distributions. It requires no approximation of a density function, which usually caused additional uncertainties [51]. Based on the original polynomial chaos expansion introduced by [82], the aPC constructs surrogate models with the help of an orthonormal polynomial basis. Such a reduction of a full biogeochemical model into a surrogate model offers the path to perform a rigorous stochastic analysis at strongly reduced computational cost.

1.4 Two-stage Bayesian model selection procedure

Bayesian model selection (BMS) (e.g. [60, 79]) has been used in many fields of research to support the choice between competing models (e.g. [9, 11, 28, 43, 57, 68, 81]). It ranks models based on their suitability to represent the available measurement data. To be more specific, BMS employs the Bayesian model evidence (BME) as the score indicating the quality of the model against the available data.

The BME-based ranking follows the principle of parsimony [67] or rather "Occam's razor", which tells us to "choose the simplest one between competing hypotheses" [31], i.e. the simplest model that can still fit the data. This results in finding the optimal trade-off between goodness-of-fit and simplicity. The work by [68] uses BME to find a justifiable level of complexity (i.e. variability of the model) for modeling a certain quantity of interest. Please note that the term "model complexity" is not uniquely defined [2, 23]. In the current study, we use complexity in the sense of "number of processes explicitly included", which is the most commonly accepted in the geoscientific community [2].

Following the framework introduced by [68], we will adopt a two-stage approach for model testing. In the first stage, the classical BMS procedure is used, in which models are tested against measurement data. This procedure is complemented by the second stage, the so-called Bayesian model justifiability analysis. Here, competing models are tested against each other based on a "synthetic truth" instead of measurement data. Based on this analysis, one can diagnose similarities between competing models and identify a suitable model that is "affordable" when only a realistic amount of measurement data is available. A joint interpretation of both stages provides insights that help find the most appropriate model, representing the observed system best under acceptable computational cost.

In the current study, we consider several models describing biogeochemical processes in subsurface porous media. They contain various assumptions helping to simplify the modeling procedure. As these models are computationally expensive, we cannot directly apply the two-stage Bayesian model selection as introduced by [68]. Instead, we base this analysis on surrogate models.

1.5 Goals and structure

The overall aim of this study is to set up a rigorous ranking of biogeochemical computationally expensive models introducing the surrogate-based two-stage Bayesian model selection procedure. We extend the Bayesian model justifiability analysis introduced by [68]. Our novel correction factor allows the use of surrogate models, making this analysis suitable for computationally demanding models.

Section 2 introduces necessary details on Bayesian updating of the aPC expansion and extends the Bayesian model selection of computationally demanding models to the Bayesian model justifiability analysis introducing novel correction factors. Section 3 introduces the biogeochemical process of microbially induced calcite precipitation (MICP) and the corresponding model set. Section 4 performs Bayesian model selection among MICP models and assesses their similarity using the novel surrogate-based justifiability analysis. Section 5 summarizes the results and gives an outlook for further investigation.

2 Bayesian assessment of computationally demanding models

2.1 Arbitrary polynomial chaos expansion

We will consider computationally demanding models, for which a straightforward application of the Bayesian model selection procedure is infeasible. Therefore, we will construct so-called surrogate models with negligible computational cost to replicate the behavior of the original physical models via the polynomial chaos expansion (PCE). The goal of PCE techniques is to construct a so-called response surface, where the modeling parameters are mapped to the model output, capturing the main features of the underlying physical model. This response surface is constructed with the help of an orthonormal polynomial basis, which is created by the Gram-Schmidt orthogonalization process [66]. Originally, it was only possible to employ this method for models with normally distributed model parameters [85]. With a generalized form, called generalized polynomial chaos (gPC) [84], the number of possible distributions for the model parameters was increased, but still limited [52]. The problem with some models is that for many model parameters the exact distribution is not known or no unique form of the distributions can be determined. Therefore, the gPC for arbitrary distributions was generalized to arbitrary polynomial chaos expansion (aPC), covering a wider range of distributions in [52]. The distributions can be discrete, continuous or discretized, they do not have to follow a certain form and can be available analytically as density function or simply as a set of samples. In this study, we use aPC to keep the proposed framework general so that it can be used for different parameter distributions. In what follows, we present the core idea for the construction of these aPC-based surrogate models.

Let $\boldsymbol{\omega} = (\omega_1, ..., \omega_{N_p})$ represent the N_p-dimensional vector of model parameters with corresponding parameter space $\boldsymbol{\Omega} = \Omega_1 \times ... \times \Omega_{N_p}$. All parameters in $\boldsymbol{\omega}$ are assumed to be independent in their prior distribution [52]. Let the model responses be given in the form of $M = f(\boldsymbol{x}, t; \boldsymbol{\omega})$, where f can be some differential equation, a coupled system of differential equations or just a simple function. Moreover, the model parameters can depend on a certain point in space $\boldsymbol{x} = (x_1, x_2, x_3)$ and time t. The model response M can be approximated with a spectral projection of responses onto orthogonal polynomial bases as follows:

$$M(\boldsymbol{x}, t; \boldsymbol{\omega}) \approx \tilde{M}(\boldsymbol{x}, t; \boldsymbol{\omega}) = \sum_{s=1}^{D} c_s(\boldsymbol{x}, t) \cdot \Psi_s(\boldsymbol{\omega}), \tag{1}$$

with the corresponding surrogate model $\tilde{M}(\boldsymbol{x}, t; \boldsymbol{\omega})$ and polynomials $\Psi_s(\boldsymbol{\omega})$ of the multivariate orthogonal polynomial basis. These polynomials are constructed according to [51]. There are D polynomials needed for the expansion, whereby D is the number of expansion coefficients dependent on the number of model parameters N_p and the chosen maximum polynomial degree d: $D = (N_p + d)!/(N_p!d!)$. The coefficients $c_s(\boldsymbol{x}, t)$ depend on space and time since the original model output depends on space and time.

To compute the coefficients $c_s(\boldsymbol{x}, t)$ of the polynomial chaos expansion in Eq. 1, we employ a non-intrusive stochastic collocation method [52]. The non-intrusiveness of this method implies that the model M can be considered as a black box, so that there is no need of modifying the governing equations of the original model at hand. Alternatively, an intrusive method such as the stochastic Galerkin method could also be used. However, as it is an intrusive method, it is necessary to modify the governing equations in the model, which can be complex [51]. Using the stochastic collocation method, a finite number of model evaluations D is sufficient to determine the coefficients. The coefficients can be computed using D evaluations of the original model M on D so-called collocation points $\left\{\omega_1^{(i)}, ..., \omega_{N_p}^{(i)}\right\}$, $i = 1, ..., D$. We solve the resulting system of equations with the help of the pseudoinverse:

$$\begin{bmatrix} \Psi_1\left(\boldsymbol{\omega}^{(1)}\right) & \dots & \Psi_D\left(\boldsymbol{\omega}^{(1)}\right) \\ \dots & \dots & \dots \\ \Psi_1\left(\boldsymbol{\omega}^{(D)}\right) & \dots & \Psi_D\left(\boldsymbol{\omega}^{(D)}\right) \end{bmatrix} \cdot \begin{bmatrix} c_1(\boldsymbol{x}, t) \\ \dots \\ c_D(\boldsymbol{x}, t) \end{bmatrix} = \begin{bmatrix} M\left(\boldsymbol{x}, t; \boldsymbol{\omega}^{(1)}\right) \\ \dots \\ M\left(\boldsymbol{x}, t; \boldsymbol{\omega}^{(D)}\right) \end{bmatrix} \tag{2}$$

or

$$\boldsymbol{\Psi}(\boldsymbol{\omega}) \cdot \mathbf{c}(\boldsymbol{x}, t) = \boldsymbol{M}(\boldsymbol{x}, t; \boldsymbol{\omega}). \tag{3}$$

The $D \times D$ matrix $\boldsymbol{\Psi}$ contains the basis polynomials, evaluated on different collocation points. The vector $\mathbf{c}$ of size $D \times 1$ contains the expansion coefficients. The outputs of the model M on the different collocation points are represented by vector $\boldsymbol{M}$ of size $D \times 1$. If one aims to compute the surrogate model of M for different points in time, it is sufficient to compute the matrix $\boldsymbol{\Psi}$ once for a fixed amount of parameters and collocation points and an expansion degree d, since the matrix is space and time independent, unlike both of the vectors $\mathbf{c}$ and $\boldsymbol{M}$. Accordingly, the coefficients are computed based on the model output using the collocation points for different points in space and time separately (Matlab code available in [49]).

The solution of the system of Eq. 3 is obviously dependent on the choice of the collocation points $\left\{\omega_1^{(i)}, ..., \omega_{N_p}^{(i)}\right\}$, $i = 1, ..., D$. According to [77] the optimal collocation points are the roots of the univariate polynomials used for the construction of the multivariate polynomial basis of degree $d + 1$ [52].

Hence, the resulting surrogate model represents the original model at the collocation points exactly while some “polynomial interpolation” is applied between them or rather an extrapolation outside of the range of the collocation points [43].

2.2 Bayesian updating of the aPC-based surrogate representation

The procedure described in Section 2.1 can be seen as an initial step, whereby the surrogate representation of the original model makes use of the prior distribution of the modeling parameters and omits the available measurement data. Therefore, the constructed surrogate model $\tilde{M}$ could be imprecise and may not necessarily cover well the region of the parameter space where the measurement data are relevant (i.e. posterior). Using a higher expansion degree to improve the surrogate model globally would increase the computational time excessively.

Therefore, to overcome this issue, we employ an iterative Bayesian updating process of the aPC representation (BaPC) that improves the accuracy of the surrogate by incorporating new collocation points at approximate locations of the maximum a posteriori parameter set [53]. The idea is to evaluate the surrogate model $\tilde{M}$ on a high number of parameter realizations, obtained from their prior distribution, to weigh the points by their posterior probability. As the parameter realization with the highest posterior probability is assumed to be in the parameter region of interest, the surrogate model should be refined there. According to the BaPC strategy, we will evaluate the original model $M(\boldsymbol{x}, t; \boldsymbol{\omega})$ on the suggested new collocation point $\boldsymbol{\omega}$ corresponding to the maximum a posteriori parameter set and recalculate the expansion coefficients $\mathbf{c}(\boldsymbol{x}, t)$ by solving Eq. 3. The increasing number of collocation points leads to an overdetermined system of equations for the determination of the coefficients which can be solved as described in Appendix A. In this way, we iteratively update the aPC representation in Eq. 1 by incorporating the points where the probability to capture the measurement data is higher. This process is repeated until the surrogate model captures the measurement data sufficiently well, although the number of iterations should be limited to keep the computational cost manageable (Matlab code available in [50]).

The suggested BaPC framework has shown promising results for computationally demanding models (e.g.

[6, 43, 54]) and further details are shown in [53]. Alternatively, other Bayesian strategies can be found in [55].

2.3 Approximation quality of aPC-based surrogate models

To assess the quality of a constructed surrogate model during the iterative Bayesian updating of an aPC expansion, we will estimate the approximation error in equation (1). Since the stochastic collocation belongs to the family of regression methods, only calculating the error at the collocation points would lead to biased results. Yet, computing the validation error via so-called testing parameter sets to assess the accuracy of the model, trained on the training collocation points, is computationally infeasible.

To remedy this problem, one can use the leave-one-out cross validation (LOOCV) as described in [7] instead. The collocation points are divided P times into two subsets, assuming that the set of collocation points is of size $P \geq D + 1$: for the calculation of the coefficients the collocation points are omitted one after the other. After the coefficients have been determined with the help of the remaining collocation points, the resulting surrogate model is evaluated on the omitted collocation point. Then, the difference to M, evaluated on this point, is computed [7]. This is done for all collocation points and finally the mean value over all quadratic errors is taken:

$$\overline{err}_{\mathrm{LOOCV}} = \frac{1}{P} \cdot \sum_{i=1}^{P} \left(M\left(\boldsymbol{\omega}^{(i)}\right) - \tilde{M}_{\backslash \boldsymbol{\omega}^{(i)}}\left(\boldsymbol{\omega}^{(i)}\right)\right)^2, \tag{4}$$

where P is the current number of collocation points, $M\left(\boldsymbol{\omega}^{(i)}\right)$ is the model evaluated on the omitted collocation point $\boldsymbol{\omega}^{(i)}$ and $\tilde{M}_{\backslash \boldsymbol{\omega}^{(i)}}\left(\boldsymbol{\omega}^{(i)}\right)$ is the surrogate model constructed without the collocation point $\boldsymbol{\omega}^{(i)}$ evaluated on the collocation point $\boldsymbol{\omega}^{(i)}$.

2.4 Bayesian model selection

Bayesian Model Selection allows to rank N_m different models M_k ($k = 1, ..., N_m$) with corresponding parameter spaces $\boldsymbol{\Omega}_k$, based on their probability to be the data-generating process (e.g. [24, 60, 79]). For this ranking, prior model weights $P\left(M_k\right)$ are updated to posterior model weights $P(M_k|\boldsymbol{y}_0)$ using Bayes' theorem:

$$P(M_k|\boldsymbol{y}_0) = \frac{p(\boldsymbol{y}_0|M_k)P(M_k)}{\sum_{i=1}^{N_{\mathrm{m}}} p(\boldsymbol{y}_0|M_i)P(M_i)}, \tag{5}$$

with $\boldsymbol{y}_0$ being the vector of measurements and the models' prior probability $P(M_k)$. The prior probability $P(M_k)$ is a subjective estimation of the investigator or the modeler about which model is an exact representation of the data-generating process, without actually knowing the data yet [60]. Uniformly distributed priors $P(M_k) = \frac{1}{N_{\mathrm{m}}}$ with N_{m} competing models are a common choice. The term $p(\boldsymbol{y}_0|M_k)$ is the so-called Bayesian Model Evidence (BME). The BME value is also known as marginal likelihood, because it can be calculated by averaging (marginalizing) over the parameter space $\boldsymbol{\Omega}_k$ of each model [32, 67]. The marginalization makes BME independent of the parameter choice and hence it is a characteristic of only the model M_k. Accordingly, BME is defined as

$$p(\boldsymbol{y}_0|M_k) = \int_{\boldsymbol{\Omega}_k} p(\boldsymbol{y}_0|M_k, \boldsymbol{\omega})\; p(\boldsymbol{\omega}|M_k)\; d\boldsymbol{\omega}, \tag{6}$$

where $p(\boldsymbol{\omega}|M_k)$ is the model-specific prior distribution of the model parameter vector $\boldsymbol{\omega} \in \boldsymbol{\Omega}_k = \Omega_1 \times ... \times \Omega_{N_p}$. The likelihood function $p(\boldsymbol{y}_0|M_k, \boldsymbol{\omega})$ quantifies how well the predictions $\boldsymbol{y}_k$ of model M_k fit the measurement data $\boldsymbol{y}_0$ and includes assumptions on the measurement error [60]. Here, we will choose a Gaussian likelihood function with zero mean:

$$p(\boldsymbol{y}_0|M_k, \boldsymbol{\omega}) = (2\pi)^{-N_{\mathrm{s}}/2}|\boldsymbol{R}|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{y}_0 - \boldsymbol{y}_k(\boldsymbol{\omega}))^T \boldsymbol{R}^{-1}(\boldsymbol{y}_0 - \boldsymbol{y}_k(\boldsymbol{\omega}))\right), \tag{7}$$

where $\boldsymbol{R}$ is the covariance matrix of the measurement error ϵ of size $N_{\mathrm{s}} \times N_{\mathrm{s}}$ (with data set size N_{s}), and $\boldsymbol{y}_k(\boldsymbol{\omega})$ is the prediction made by model M_k with the model parameter vector $\boldsymbol{\omega}$.

For most applications, there is no analytical solution of Eq. 6 and the corresponding integral can be estimated using a brute-force Monte Carlo approach, which yields an unbiased approximation. To perform the Monte Carlo integration, we create a sample set of N_{MC} realizations of the modeling parameter vector $\boldsymbol{\omega}$ based on its prior distribution $p(\boldsymbol{\omega}|M_k)$. With the corresponding likelihood functions Eq. 7, we will obtain the following numerical approximation of the BME value:

$$p(\boldsymbol{y}_0|M_k) \approx \frac{1}{N_{\mathrm{MC}}} \sum_{i=1}^{N_{\mathrm{MC}}} p(\boldsymbol{y}_0|M_k, \boldsymbol{\omega}_i), \tag{8}$$

where $\boldsymbol{\omega}_i$ is the i-th parameter realization for model M_k.

2.5 aPC-based Bayesian model selection

Remarking that the surrogate representation $\tilde{M}_k$ is only an approximation of the original model M_k, we expect that surrogate-based BME values could be misleading for the Bayesian model selection procedure. Therefore, conclusions drawn from BME values based on surrogates are only valid to the degree of the approximation quality of the surrogate model. Such falsified values can be avoided by adapting the calculation of the BME value, as proposed in [43]. We will consider that the prediction of the surrogate model $\tilde{M}_k$ contains an approximation error E_k. We consider it to be independent of the measurement error ϵ (because E_k and ϵ have no interaction), so that $M_k = \tilde{M}_k + E_k$. Therefore $p(\boldsymbol{y}_0|\tilde{M}_k + E_k, \boldsymbol{\omega}) = p(\boldsymbol{y}_0|\tilde{M}_k, \boldsymbol{\omega}) \cdot p(M_k|\tilde{M}_k, \boldsymbol{\omega})$ and the BME value in Eq. 6 can be rewritten as:

$$p(\boldsymbol{y}_0|M_k) = \int_{\boldsymbol{\Omega}_k} p(\boldsymbol{y}_0|\tilde{M}_k, \boldsymbol{\omega})\ p(M_k|\tilde{M}_k, \boldsymbol{\omega})\ p(\boldsymbol{\omega}|M_k)\ d\boldsymbol{\omega}, \tag{9}$$

where $p(M_k|\tilde{M}_k, \boldsymbol{\omega})$ is the likelihood function that indicates how well the original model prediction based on the model parameter realization $\boldsymbol{\omega}$ matches the corresponding surrogate model prediction:

$$\begin{aligned} p(M_k|\tilde{M}_k, \boldsymbol{\omega}) &= (2\pi)^{-N_s/2}|\boldsymbol{S}|^{-1/2} \\ &\cdot \exp\left(-\frac{1}{2}(\boldsymbol{y}_k(\boldsymbol{\omega}) - \tilde{\boldsymbol{y}}_k(\boldsymbol{\omega}))^T \boldsymbol{S}^{-1}(\boldsymbol{y}_k(\boldsymbol{\omega}) - \tilde{\boldsymbol{y}}_k(\boldsymbol{\omega}))\right), \end{aligned} \tag{10}$$

with the predictions $\boldsymbol{y}_k$ of the original model M_k and $\tilde{\boldsymbol{y}}_k$ of the surrogate model $\tilde{M}_k$ and the covariance matrix $\boldsymbol{S}$ of approximation errors.

Following the derivation in [43], we obtain the corrected BME value for the original model, computed on the basis of the reduced model:

$$p(\boldsymbol{y}_0|M_k) = p(\boldsymbol{y}_0|\tilde{M}_k) \cdot \int_{\boldsymbol{\Omega}_k} p(M_k|\tilde{M}_k, \boldsymbol{\omega})\ p(\boldsymbol{\omega}|\tilde{M}_k, \boldsymbol{y}_0)\ d\boldsymbol{\omega}. \tag{11}$$

Equation 11 shows clearly how the BME value of the original model (BME_{OM}) can be calculated from the BME value of the surrogate model (BME_{SM}):

$$\text{BME}_{\text{OM}} = \text{BME}_{\text{SM}} \cdot \text{Weight}_{\text{SM}}, \tag{12}$$

with

$$\begin{aligned} \text{BME}_{\text{OM}} &= p(\boldsymbol{y}_0|M_k), \\ \text{BME}_{\text{SM}} &= p(\boldsymbol{y}_0|\tilde{M}_k) \text{ and} \\ \text{Weight}_{\text{SM}} &= \int_{\boldsymbol{\Omega}_k} p(M_k|\tilde{M}_k, \boldsymbol{\omega})\ p(\boldsymbol{\omega}|\tilde{M}_k, \boldsymbol{y}_0)\ d\boldsymbol{\omega}, \end{aligned} \tag{13}$$

where the BME_{SM} value can be computed as described in the previous section, using the surrogate model $\tilde{M}_k$ instead of the original model M_k.

The correction factor $\text{Weight}_{\text{SM}}$ requires an integration over the whole parameter space $\boldsymbol{\Omega}_k$ and its computation via Monte Carlo Integration is not feasible due to the high computational cost of the original model. Therefore, the correction factor can be estimated at those collocation points $\boldsymbol{\omega}^*$ that were used to construct the surrogate model:

$$\text{Weight}_{\text{SM}} \approx \sum_{i=1}^{P} p(M_k|\tilde{M}_k, \boldsymbol{\omega}_i^*)\ p(\boldsymbol{\omega}_i^*|\tilde{M}_k, \boldsymbol{y}_0), \tag{14}$$

where P is the number of collocation points. Using only the collocation points to calculate the correction factor leads to the fact that $\text{BME}_{\text{SM}} \cdot \text{Weight}_{\text{SM}}$ is not equivalent to BME_{OM}, but is merely an approximation. However, the corrected BME_{SM} is a better approximation of BME_{OM} than BME_{SM} without correction [43].

2.6 Bayesian model justifiability analysis

In order to complement the comparison of the models against the measurement data, [68] suggested a so-called Bayesian model justifiability analysis, in which the competing models are tested against each other in a synthetic setup omitting the measurement data. The justifiability analysis can help to decide whether the apparently most appropriate model from the conventional BMS analysis is really the best model in the set or whether this model is only optimal given the limited amount of available measurement data [68]. Additionally, the justifiability analysis provides insights about similarities among the tested models.

To perform the justifiability analysis, we will generate the so-called model confusion matrix [68]. Confusion matrices are typically used in the field of statistical classification (e.g. [1]) to compare the actual and the predicted classification, visualizing whether an object is misclassified ("confused"). In that way, we can recognize whether a model is able to distinguish its own predictions from the ones of its competitors. To do so, we calculate the Bayesian model weights for all models adopting (5).

However, instead of using the measurement data $\boldsymbol{y}_0$, each of the competing models generates a finite series of prior predictions that serve as realizations of the "synthetic truth". Thus, we generate N_{MC} synthetic data sets of each model based on samples of its prior parameter distributions. Then, each synthetic data set is compared to the competing models by first computing the likelihood function as described in Eq. 7, for example of the single realization i of model M_k based on the data set j of model M_l. The BME value can be obtained by calculating the mean of all likelihoods $p(M_{l,j}|M_k)$ of model M_k given this single realization j of model M_l. The resulting model confusion matrix has the size $N_{\text{m}} \times N_{\text{m}}$, for N_{m} competing models.

To execute both steps of model testing ((1) BMS testing against measurements and (2) justifiability analysis testing models against each other) simultaneously, we add the measurement data to our model set, i.e. we add it as a new row and column to the confusion matrix.

A schematic illustration of its construction is given in Fig. 1, whereby the model confusion matrix is extended by the standard BMS procedure (i.e. including measurements).

The blue box in Fig. 1 represents a standard BMS procedure where the model M_k has been tested against the measurement data. This entry can be obtained from Eq. 6, using Monte Carlo Integration for $p(\boldsymbol{y}_0|M_k)$ as in Eq. 8. The green box in Fig. 1 reflects the likelihood of a single realization of model M_k given a single realization

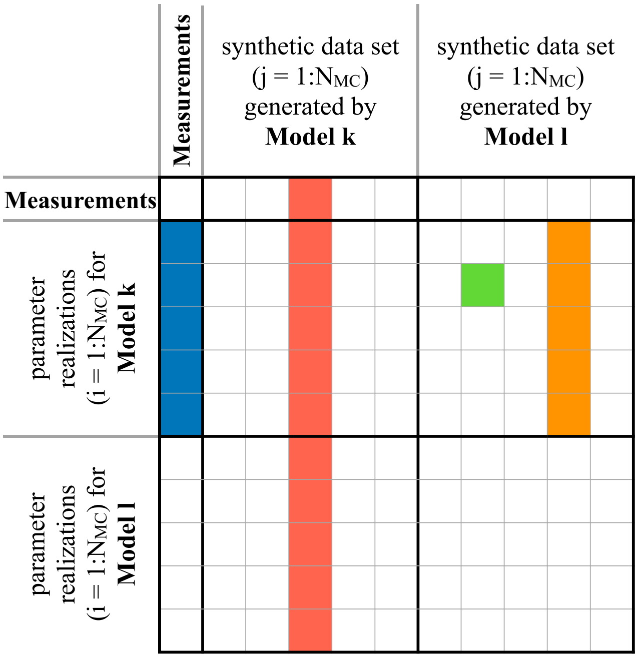


Fig. 1 Schematic illustration of constructing the model confusion matrix

of the reference model M_l, which currently serves as synthetic truth. The orange box in Fig. 1 shows the average likelihood (BME) of model M_k given a single realization of the reference model M_l. This BME value is normalized by the sum of the BME values of all models given a single realization of the synthetic truth (red box), yielding a posterior model weight $p(M_l|M_{k,j})$ with the reference model M_k. The bold boxes in Fig. 1 illustrate these averaged posterior weights over all synthetic data sets of the reference model M_k. The bold boxes of one column contain the expected posterior weights (PW) of all models given that model M_k is true. One entry can be computed as follows:

$$PW_{lk} = \frac{1}{N_{\mathrm{MC}}} \sum_{j=1}^{N_{\mathrm{MC}}} p(M_l|M_{k,j}) \tag{15}$$

$$= \frac{1}{N_{\mathrm{MC}}^2} \sum_{j=1}^{N_{\mathrm{MC}}} \sum_{i=1}^{N_{\mathrm{MC}}} p(M_{l,i}|M_{k,j}), \tag{16}$$

whereby the averaged BME value $\left(\sum_{i=1}^{N_{\mathrm{MC}}} p(M_{l,i}|M_{k,j})\right)$ in Eq. 16 is not normalized for the sake of readability.

The resulting extended model confusion matrix consists only of these entries, i.e. the bold boxes and therefore has the size $(N_{\mathrm{m}}+1) \times (N_{\mathrm{m}}+1)$ for N_{m} competing models and the measurement data.

The main diagonal entries reflect how good each model identifies itself as the data-generating process, given a certain data set size. The values of the diagonal entries should be equal to 1.00 with an infinite data set size. However, for finite data sets, models might “confuse” their own predictions (misclassification) with the ones of competing models due the two following reasons. (1) Two models are actually highly similar. (2) One model has a high goodness-of-fit to the reference data, but also a high variability in its predictions. The BMS framework punishes this high variability with a lower model weight. Thus, a scenario of a less variable model, which fits the reference data worse than the more variable one, might lead to similar model weights. When more synthetic data is used, the more variable model will receive a higher weight, as its variability becomes more justifiable, while the weight of the less variable model will decrease [23, 24].

The off-diagonal entries of the model confusion matrix reflect the similarity between pairs of models. This can be useful when comparing possible simplifications to a detailed reference model [65]. With the aid of the model confusion matrix it is possible to identify the model that yields the most similar results to the reference model at reduced computational cost.

2.7 aPC-based Bayesian model justifiability analysis

We will combine the methodologies from Sections 2.5 and 2.6 towards an aPC-based Bayesian model justifiability analysis, where models are mutually tested against each other. To do so, we will consider two models, model M_k and model M_l. The comparison of two models implies that one model, M_l in this case, is assumed to be the data-generating process. Instead of computing the BME value for the original models $p(M_l|M_k)$, we have to calculate the BME value $p(\tilde{M}_l|\tilde{M}_k)$ of the surrogate models. Similar to Section 2.5, we assume that each surrogate representation of each analyzed model contains an approximation error: $M_k = \tilde{M}_k + E_k$ and $M_l = \tilde{M}_l + E_l$. Therefore, Eq. 11 can be rewritten as:

$$p(M_l|M_k) = p(M_l|\tilde{M}_k) \cdot \int_{\boldsymbol{\Omega}_k} p(M_k|\tilde{M}_k, \boldsymbol{\omega})\, p(\boldsymbol{\omega}|\tilde{M}_k, M_l)\, d\boldsymbol{\omega}. \tag{17}$$

In the next step, we focus on the term $p(M_l|\tilde{M}_k)$, considering $M_l = \tilde{M}_l + E_l$ leads us to

$$p(M_l|\tilde{M}_k) = \int_{\boldsymbol{\Omega}_k} p(\tilde{M}_l|\tilde{M}_k, \boldsymbol{\omega})\, p(M_l|\tilde{M}_l, \boldsymbol{\omega}_k)\, p(\boldsymbol{\omega}|\tilde{M}_k)\, d\boldsymbol{\omega}. \tag{18}$$

Multiplying and dividing the right-hand side of Eq. 18 by $p(\tilde{M}_l|\tilde{M}_k)$ and applying Bayes’ theorem yields

$$p(M_l|\tilde{M}_k) = p(\tilde{M}_l|\tilde{M}_k) \cdot \int_{\boldsymbol{\Omega}_k} p(M_l|\tilde{M}_l, \boldsymbol{\omega})\, p(\boldsymbol{\omega}|\tilde{M}_k, \tilde{M}_l)\, d\boldsymbol{\omega}. \tag{19}$$

When inserting Eq. 19 into Eq. 17, we obtain

$$\begin{aligned} p(M_l|M_k) &= p(\tilde{M}_l|\tilde{M}_k) \\ &\cdot \int_{\boldsymbol{\Omega}_k} p(M_l|\tilde{M}_l,\boldsymbol{\omega})p(\boldsymbol{\omega}|\tilde{M}_k,\tilde{M}_l)\,d\boldsymbol{\omega} \\ &\cdot \int_{\boldsymbol{\Omega}_k} p(M_k|\tilde{M}_k,\boldsymbol{\omega})\;p(\boldsymbol{\omega}|\tilde{M}_k,M_l)\,d\boldsymbol{\omega}, \end{aligned} \tag{20}$$

or

$$\mathrm{BME}_{\mathrm{OMOM}} = \mathrm{BME}_{\mathrm{SMSM}} \cdot \mathrm{Weight}_{\mathrm{SM1}} \cdot \mathrm{Weight}_{\mathrm{SM2}}, \tag{21}$$

with

$$\begin{aligned} \mathrm{BME}_{\mathrm{OMOM}} &= p(M_l|M_k) \\ \mathrm{BME}_{\mathrm{SMSM}} &= p(\tilde{M}_l|\tilde{M}_k) \\ \mathrm{Weight}_{\mathrm{SM1}} &= \int_{\boldsymbol{\Omega}_k} p(M_l|\tilde{M}_l,\boldsymbol{\omega})\;p(\boldsymbol{\omega}|\tilde{M}_k,\tilde{M}_l)\,d\boldsymbol{\omega} \\ \mathrm{Weight}_{\mathrm{SM2}} &= \int_{\boldsymbol{\Omega}_k} p(M_k|\tilde{M}_k,\boldsymbol{\omega})\;p(\boldsymbol{\omega}|\tilde{M}_k,M_l)\,d\boldsymbol{\omega}, \end{aligned} \tag{22}$$

whereby $\mathrm{BME}_{\mathrm{OMOM}}$ corresponds to the BME value when comparing two original models and $\mathrm{BME}_{\mathrm{SMSM}}$ to the BME value when comparing two surrogate models. The value of $\mathrm{BME}_{\mathrm{SMSM}}$ can be computed in the same way as proposed in Eq. 6 via Monte Carlo integration in Eq. 8 with the likelihood function defined in Eq. 7, using the prediction of model M_l evaluated on a certain model parameter vector $\boldsymbol{\omega}$ instead of the measurement data $\boldsymbol{y}_0$. The collocation points $\boldsymbol{\omega}^*$ can be employed again similarly to Section 2.5 to compute the correction factors for both models:

$$\begin{aligned} \mathrm{Weight}_{\mathrm{SM1}} &\approx \sum_{i=1}^{P} p(M_l|\tilde{M}_l,\boldsymbol{\omega}_i^*)\;p(\boldsymbol{\omega}_i^*|\tilde{M}_k,\tilde{M}_l) \\ \mathrm{Weight}_{\mathrm{SM2}} &\approx \sum_{i=1}^{P} p(M_k|\tilde{M}_k,\boldsymbol{\omega}_i^*)\;p(\boldsymbol{\omega}_i^*|\tilde{M}_k,M_l). \end{aligned} \tag{23}$$

Moreover, since the model confusion matrix in the Bayesian model justifiability framework compares the original models as well, we have to account for the approximation of these models with the surrogates. As the weights $\mathrm{Weight}_{\mathrm{SM1}}$ and $\mathrm{Weight}_{\mathrm{SM2}}$ are not dependent on a single parameter realization, the overall posterior weights of the model confusion matrix can be corrected in the same way as the BME values. To this end, the posterior values (PW) of the model confusion matrix from Eq. 16 need to be multiplied by the two correction factors $\mathrm{Weight}_{\mathrm{SM1}}$ and $\mathrm{Weight}_{\mathrm{SM2}}$ from Eq. 23:

$$\begin{aligned} PW_{lk} &= \frac{1}{N_{\mathrm{MC}}}\sum_{j=1}^{N_{\mathrm{MC}}} p(M_l|M_{k,j}) \\ &= \frac{1}{N_{\mathrm{MC}}}\sum_{j=1}^{N_{\mathrm{MC}}} p(\tilde{M}_l|\tilde{M}_{k,j}) \cdot \mathrm{Weight}_{\mathrm{SM1}} \cdot \mathrm{Weight}_{\mathrm{SM2}}, \end{aligned} \tag{24}$$

where $\mathrm{SM1} = \tilde{M}_l$ and $\mathrm{SM2} = \tilde{M}_k$.

3 Biogeochemical processes in porous media

3.1 Microbially induced calcite precipitation

Microbially induced calcite precipitation (MICP) is a typical biogeochemical process. When conceptualizing MICP in porous media, various phases are involved: there are at least three solid phases (biofilm, calcite and unreactive solid material), water and possibly another fluid phase, e.g. gas. Additionally, at least calcium, inorganic carbon, and urea are considered as dissolved components in the water phase, the complete list of components can be found in [25].

MICP is a reactive transport process consisting of three main parts: (1) adhesion of biomass on surfaces, detachment of the biomass from the biofilm as well as growth and decay of the biomass, (2) urea hydrolysis that alters the geochemistry and (3) precipitation and dissolution of calcite. A visualization of the MICP process is shown in Fig. 2.

S. pasteurii are bacteria that are able to produce the enzyme urease and to decompose urea into carbonic acid and ammonia with the aid of urease. In aqueous solution, the ammonia reacts with the contained H^+ ions. As a result, the pH value increases so that the carbonic acid decomposes into H^+ ions and carbonate ions, while the concentration of dissolved carbonate increases. If calcium ions are provided, it comes to a reaction with the carbonate ions and calcite precipitates.

Shortly, all together this leads to the following MICP reaction equation [25]:

$$CO(NH_2)_2 + 2H_2O + Ca^{2+} \xrightarrow{Urease} 2NH_4^+ + CaCO_3 \downarrow . \tag{R1}$$

3.2 Experimental setup

The analyzed MICP experiment is described in detail in [25] (there, see experiment "D1"). It describes a sand-filled column that is 61 cm high with a diameter of 2.54 cm. In the beginning of the experiment, bacteria are injected at the bottom of the column. Bacteria are allowed to attach

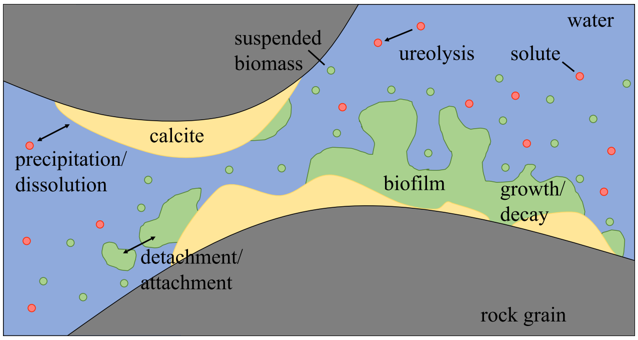


Fig. 2 Schematic view of relevant processes and phases during MICP after [25]

during an over-night no-flow period establishing a biofilm throughout the column. Then, biofilm growth is promoted by a 24 hour substrate injection. From there, two pore volumes of 0.33 mol/l calcium and urea solution are injected at 10 ml/min repeatedly every 24 hours. The no-flow period after the injection allows the mineralization reactions to take place. That period is followed by another injection of substrate to revive the biofilm [25], before the next injections of calcium and urea start over until a total number of 30 cycles was reached. A schematic experiment setup is shown in Fig. 3.

For this analysis, out of various predicted quantities we pick only the model predictions of calcium and calcite over space and time. The predictions of different models are compared to measurement data as well as among each other. In order to receive comparable results, only spatial and temporal points where measurement data are available are used when comparing models among each other. These data points differ for calcium and calcite. For the calcite content, measurement data are only available at the end of the experiment, which is after 3203460 seconds (about 890 hours or 37 days). The calcium concentration is measured at 35 different data points in time. Therefore, calcium concentrations are measured after 6 "main points" in time, the so-called pulses, namely after 151.35, 218.85, 290.85, 626.85, 698.85 and 866.85 hours. At these points, the concentration is measured and additionally after half an hour, one, two, three and four hours, except for pulse 22, where no measurement is available after 3 hours, which results in 35 temporal points. The exact times of measurement after the first injection can be taken from Table 1.

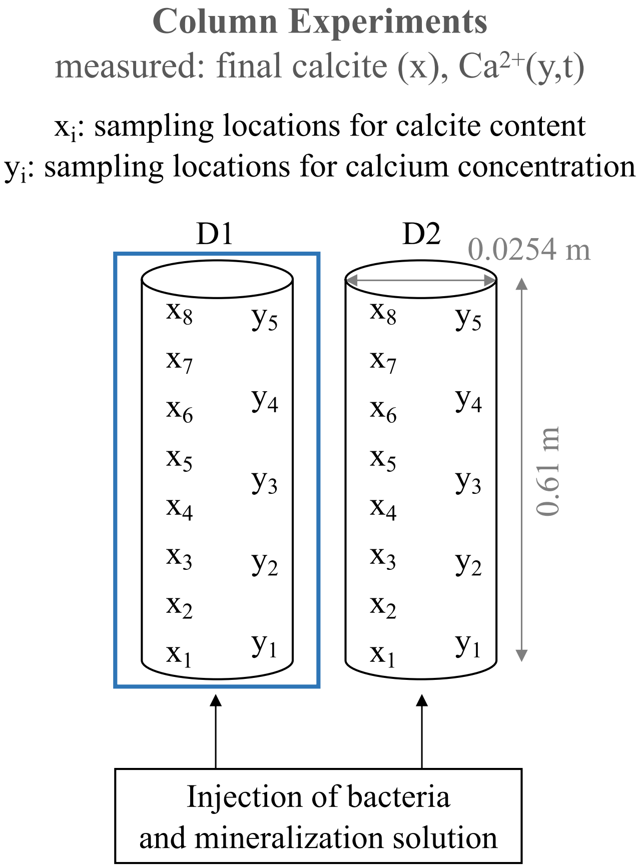

Column Experiments
measured: final calcite (x), Ca^{2+}(y,t)
x_i: sampling locations for calcite content
y_i: sampling locations for calcium concentration
D1
D2
0.0254 m
0.61 m
x_8
x_7
x_6
x_5
x_4
x_3
x_2
x_1
y_5
y_4
y_3
y_2
y_1
Injection of bacteria
and mineralization solution

Fig. 3 Column experiment setup by [25] with measurement locations for calcite content and calcium concentration with analyzed column D1

There are eight measurement locations for the calcite concentration, located at 3.81, 11.43, 19.05, 26.67, 34.29, 41.91, 49.53 and 57.15 cm distance from the bottom. For the calcium concentration, there are only five spatial measurement points located at 10.16, 20.32, 30.48, 39.37 and 49.53cm distance from the bottom. The measurement locations in the models are evenly distributed at a respective distance of half an inch (1.27 cm).

3.3 Conceptual models and related uncertainty

We analyze three models for MICP that describe biogeochemical processes in porous media provided by [25, 26]. For detailed explanation of their equations and the used numerical schemes, we refer to that original publication. All models account for changes in porosity and permeability and use the same discretization and solution strategy: a fully implicit Euler scheme in time and fully-coupled-vertex-centered finite volume (box) scheme [21] in space; the system of equations is solved using the BiCGStab solver [78] after linearization using the Newton–Raphson method.

An <Intel(R) Xeon(R) CPU E5-2680 v2 @2.80 GHz, 40 Cores> machine was used for the model evaluations. The computational effort for the most detailed MICP model, referred to as *full complexity* model, is extremely high with a run time between 16 and 42 hours, depending on the respective model parameter set. The exact cost is dependent on the model parameter set chosen for the evaluation, since the time stepping varies adaptively. Therefore, [25] suggest two simplifications of the *full complexity* model M_{FC} using the following physical assumptions.

- *initial biofilm* model (M_{IB}): The suspended biomass is neglected and the biofilm is assumed to be already established at the beginning of the experiment.
- *simple chemistry* model (M_{SC}): The ureolysis rate is the rate limiting reaction and precipitation of calcite occurs immediately whenever urea is hydrolyzed as described in the overall reaction (R1) [26].

As described in Section 3.2, the experiment starts with a biomass injection and a growth period until the biofilm is established. The *initial biofilm* model M_{IB} omits this part of the simulation under the assumption that a uniformly distributed biofilm is already established in the beginning

Table 1 Times in hours for measurement of the calcium concentration

pulse number after pulse	5	7	10	22	24	30
0 hours	151.35	218.85	290.85	626.85	698.85	866.85
0.5 hours	151.85	219.35	291.35	627.35	699.35	867.35
1 hour	152.35	219.85	291.85	627.85	699.85	867.85
2 hours	153.35	220.85	292.85	628.85	700.85	868.85
3 hours	154.35	221.85	293.85	-	701.85	869.85
4 hours	155.35	222.85	294.85	630.85	702.85	870.85

of the experiment and assuming further no reattachment of biomass detached from the biofilm. As growth, decay and detachment of biofilm are still considered, leading to non-uniform biofilm along the flow path, the initial distribution of biofilm has very limited impact on the simulation results for the used injection strategy [27]. Additionally, the number of necessary primary variables is reduced by one, as suspended biomass does not need to be considered [26]. The *simple chemistry* model M_{SC} simplifies the precipitation rate equation to be equal to the ureolysis rate equation. The model makes the assumption that whenever urea put into the system hydrolyzes, calcite immediately precipitates, treating calcite precipitation as an equilibrium reaction and ignoring the saturation state. Therefore, there is no need for computing the precipitation rate and the associated expensive-to-calculate saturation state and carbonate and calcium activities. As the activities do not need to be calculated, also the ammonia/ammonium produced during ureolysis do not have any effects on the precipitation rate and thus, the results. Consequently, the primary variable accounting for ammonia/ammonium is removed, reducing the number of primary variables by one [26]. The key differences that are important for the model simplifications are summarized in Table 2.

The computational time of the *initial biofilm* model M_{IB} still remains high and is only slightly lower than for the *full complexity* model on the same computational cluster. The strong assumptions in the *simple chemistry* model M_{SC} allow to obtain results of one model run after 40 minutes using the same computational cluster. Apart from decreasing the computational cost, model simplification reduces parametric uncertainty. A too detailed (too complex) model with many parameters and without enough calibration data and therefore parametric uncertainty results in a high predictive variance (i.e. uncertainty) of the model.

Models should generally be "as simple as possible, as complex as necessary" (principle of parsimony) [23] to prevent overfitting (e.g. [3, 36]). The considered parameters in the following were previously identified as sensitive parameters of the MICP models and already used for calibration in [25]:

- the coefficient for preferential attachment to biomass $c_{a,1}$, $\left[s^{-1}\right]$
- the coefficient for attachment to arbitrary surfaces $c_{a,2}$, $\left[s^{-1}\right]$
- the dry mass density of biofilm ρ_f, $\left[\mathrm{kg/m^3}\right]$
- the enzyme content of biomass k_{ub}, $\left[\mathrm{kg/kg}\right]$.

As the *initial biofilm* model M_{IB} assumes that there are no attachment periods, it is only dependent on the model parameters ρ_f and k_{ub}. The *full complexity* model M_{FC} and *simple chemistry* model M_{SC} are both dependent on all four model parameters. Following the physically possible range of the considered uncertain parameters, we assume that all of the model parameters are uniformly distributed in the intervals shown in Table 3.

Table 2 Key differences of the investigated models

model	full complexity M_{FC}	initial biofilm M_{IB}	simple chemistry M_{SC}
simplifying assumption	–	pre-existing biofilm	precipitation determined by ureolysis
simulated time	3203460 s	3109860 s	3203460 s
biomass transport and attachment	yes	no	yes
sophisticated geochemistry	yes	yes	no
kinetic precipitation rate	yes	yes	no
number of primary variables	12	11	11
neglected component	–	suspended biomass	ammonia/ammonium

Table 3 Intervals for the model parameters

model parameter	interval
$c_{\mathrm{a},1}$	$\left[1 \cdot 10^{10} s^{-1}, 1 \cdot 10^{-7} s^{-1}\right]$
$c_{\mathrm{a},2}$	$\left[1 \cdot 10^{10} s^{-1}, 1 \cdot 10^{-6} s^{-1}\right]$
ρ_{f}	$\left[1\ \mathrm{kg/m^3}, 15\ \mathrm{kg/m^3}\right]$
k_{ub}	$\left[1 \cdot 10^{-5}\ \mathrm{kg/kg}, 5 \cdot 10^{-4}\ \mathrm{kg/kg}\right]$

3.4 Implementation details of the surrogate models

We construct two surrogate models (one for calcite, one for calcium) for each of the three competing MICP models described in Section 3.3 (resulting in a total of six different surrogate models) using a $d = 2$ order aPC expansion according to the prior distributions presented in Table 3. For this purpose, the three original models will be evaluated $D = (N_p + d)!/(N_p!d!)$ times according to Section 2.1. Since the D evaluations for the construction of the surrogate models are independent, these model runs were parallelized. Further, we refine each of the three surrogates using iterative Bayesian updating of the aPC representation according to Section 2.2. Here, we restrict the number of Bayesian updates to ten due to the high computational demand and previous experience (see e.g. [6]), so that $P_{\mathrm{end}} = D + 10 = (N_p + d)!/(N_p!d!) + 10$. This results in $P_{\mathrm{end}} = 15 + 10 = 25$ model evaluations for the *simple chemistry* model M_{SC} and the *full complexity* model M_{FC} and $P_{\mathrm{end}} = 6 + 10 = 16$ for the *initial biofilm* model M_{IB}. During the Bayesian updating, we consider the standard deviation of measurement errors ϵ at each point in space (and time) equal to 20% of the associated measurement value for both the calcite content and the calcium concentration.

4 Bayesian model justifiability analysis of Biogeochemical models in porous media

4.1 aPC-based representation of MICP models

Equation 4 provides errors of the surrogate models for every point in space and time due to the structure of Eq. 1. As every point in space and time has its own surrogate model, there are $5 \cdot 35 \cdot 10 = 1750$ LOOCV errors (5 spatial and 35 temporal points that are used for the comparison, 10 updating steps) computed for calcium and $8 \cdot 10$ for calcite (8 spatial points that are used for the comparison, 10 updating steps) in the analyzed set up. The LOOCV error is computed after the primal construction of the surrogate models and during the iterative Bayesian updating. In order to visualize the errors, we will average the respective values over space (and time) after every updating step. In order to compare the LOOCV error of the surrogate models for calcium and calcite, the relative errors must be considered, since the two quantities of interest (calcite content [%] and calcium concentration [mol/m^3]) are in different orders of magnitude. For this purpose, they were normalized to the mean output value, as shown in Fig. 4.

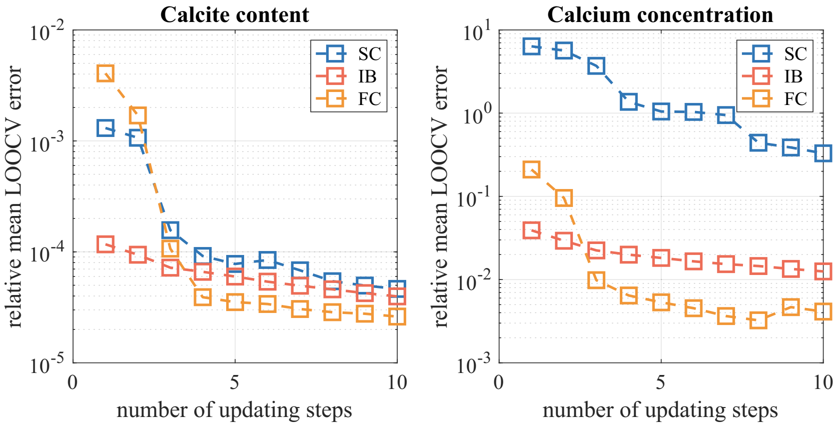
The relative mean LOOCV errors before the first update are not considered in this figure to get a better visualization, since this error is significantly higher than the ones after the updates. First of all, the figure shows that the error for calcite decreases more strongly than the error for calcium. It is also remarkable that for all models the error for calcite is in a similar order of magnitude. This means that all surrogate models are of a comparable quality for the calcite content. For calcium, the error of the *simple chemistry* model M_{SC} is significantly larger than the one for the other two surrogate models. This can occur if one uses Bayesian updating and wants to improve the models only in the region of the measurement data. This means the surrogate model is similar to the original one in the region of the measurement data, but it deviates a lot from the original model in other regions (not part of the measurement points). This results in a higher overall LOOCV error. The larger error of the surrogate model is compensated later by the newly introduced correction factor in Section 2.5.

Furthermore, the relative mean LOOCV errors for calcite are in a range of $[2 \cdot 10^{-5}, 6 \cdot 10^{-5}]$ after the last update and those for calcium are in a range of $[4 \cdot 10^{-3}, 4 \cdot 10^{-1}]$. Accordingly, the worst surrogate response for calcite is still better than the best one for calcium. This indicates that the surrogate models for the calcite content as a whole are better with respect to the LOOCV error than those for the calcium concentration.

4.2 aPC-based Bayesian model justifiability analysis for MICP models

We will perform the aPC-based Bayesian model selection incorporating the measurement data and aPC-based Bayesian model justifiability analysis according to Sections 2.5 and 2.7 using the obtained surrogate representations of the three analyzed MICP models from Section 4.1. Following the justifiability analysis, we compute the model weights as stated in Section 2.6 and adjust them with the novel correction factors from Sections 2.5 and 2.7 in a second stage. BME convergence was ensured by checking the evolution of the averaged likelihood over an increasing data set size. In order to justify the underlying physical assumptions behind the MICP models, we will assess the impact of the data set size onto BME values appearing in the Bayesian model justifiability analysis. To do so, we start with only one spatial data point, then we use half of the available data set size and finally we include all of the spatial data points for calcium and calcite. This results in the

Fig. 4 Relative mean LOOCV errors for calcite content and calcium concentration with increasing number of updates



following data set sizes $N_{D,\text{spatial}} \in \{1, 3, 5\}$ for calcium and $N_{D,\text{spatial}} \in \{1, 4, 8\}$ for calcite.

4.2.1 aPC-based BMS and Bayesian model justifiability analysis

In a first stage, the conventional BMS analysis for measurement data is performed with results illustrated in Fig. 5.

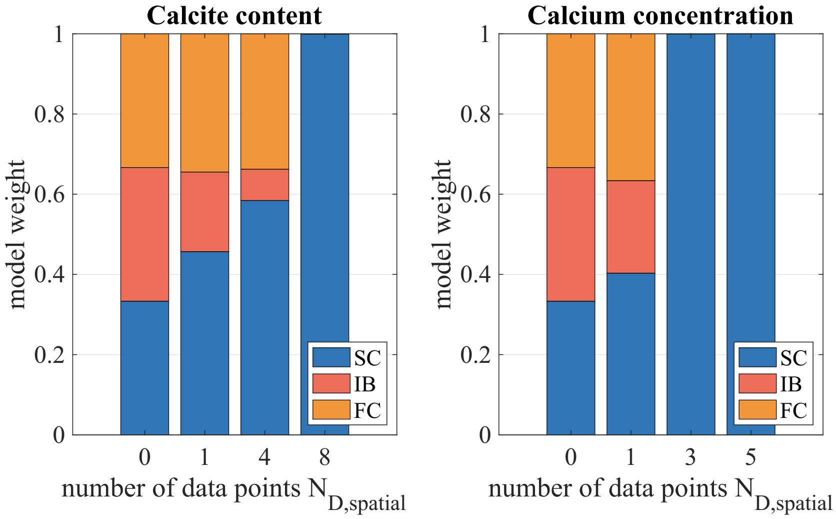
One can observe that the *simple chemistry* model M_{SC} obtains the highest model weight (normalized BME value) for all data set sizes. A model wins the competition either because of its low complexity or because of its goodness-of-fit to the measurement data (or both) [68]. These two aspects will be further investigated in a second stage, the justifiability analysis.

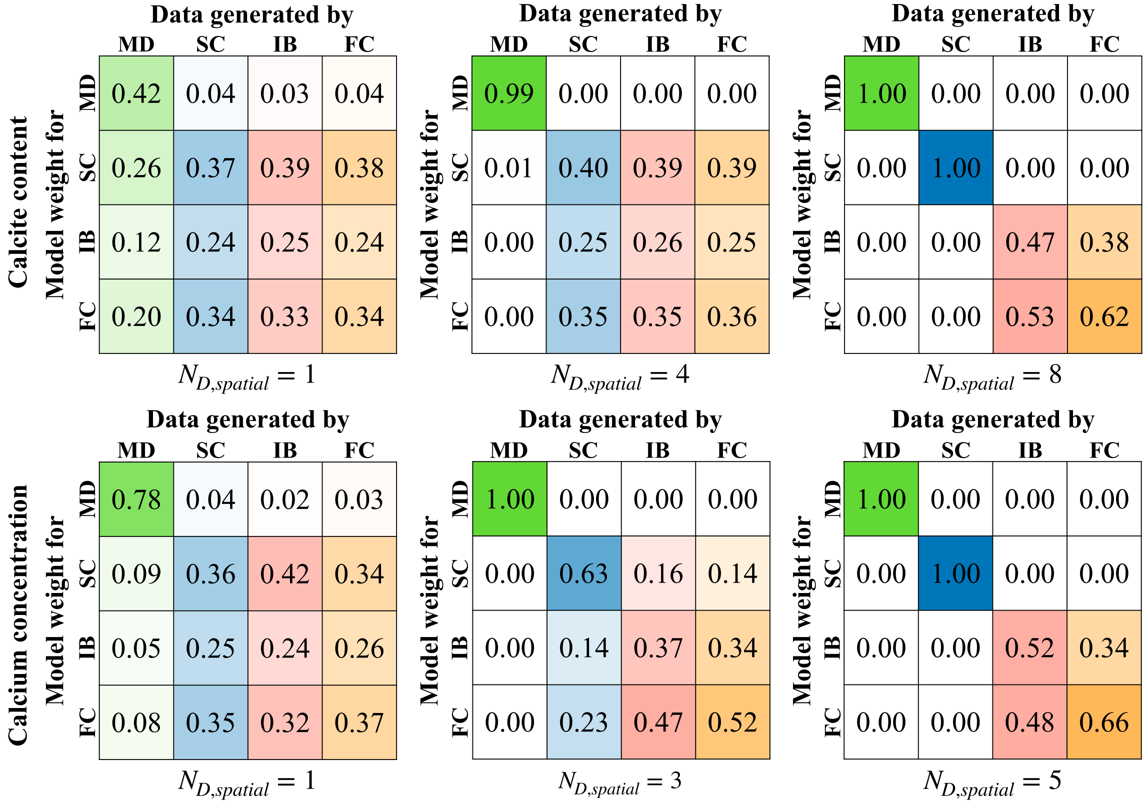
Figure 6 shows the corresponding model confusion matrices for both the calcite content and the calcium concentration predictions. Each entry corresponds to the weight of one model, which is the probability that model M_k (rows) is the data-generating process of the predictions made by model M_l (columns) according to Bayes' theorem.

The main-diagonal entries of the model confusion matrices in Fig. 6 represent the models' ability to identify their own predictions. The higher the value of the main diagonal entry in Fig. 6, the higher is the probability of the model to identify itself as the data-generating process. The diagonal values increase when a bigger data set size is used, agreeing well with the theory of the Bayesian model justifiability analysis discussed in [68]. The diagonal weight of the simplest model, the *simple chemistry* model M_{SC}, is always the highest, independent of the data set size, which shows that the analysis identifies this model as data-generating, even if the data set is large and the model makes strong assumptions. For both the calcium and the calcite, the diagonal entries achieve the "absolute majority" of more than 0.50 in favor of justifiability (except for the *initial biofilm* model M_{IB} for calcite) when taking the full data set into account. This means that the data set size is sufficient to justify the modeling concepts behind the considered models.

But even for the full data set, the *full complexity* model M_{FC} obtains a high weight when the *initial biofilm* model M_{IB} generates the data and vice versa. It follows that the

Fig. 5 Model weights for the prediction of calcite content and calcium concentration over increasing amount of used spatial data points $N_{D,\text{spatial}}$





Calcite content

Model weight for / Data generated by

	MD	SC	IB	FC
MD	0.42	0.04	0.03	0.04
SC	0.26	0.37	0.39	0.38
IB	0.12	0.24	0.25	0.24
FC	0.20	0.34	0.33	0.34

$N_{D,spatial} = 1$

Model weight for / Data generated by

	MD	SC	IB	FC
MD	0.99	0.00	0.00	0.00
SC	0.01	0.40	0.39	0.39
IB	0.00	0.25	0.26	0.25
FC	0.00	0.35	0.35	0.36

$N_{D,spatial} = 4$

Model weight for / Data generated by

	MD	SC	IB	FC
MD	1.00	0.00	0.00	0.00
SC	0.00	1.00	0.00	0.00
IB	0.00	0.00	0.47	0.38
FC	0.00	0.00	0.53	0.62

$N_{D,spatial} = 8$

Calcium concentration

Model weight for / Data generated by

	MD	SC	IB	FC
MD	0.78	0.04	0.02	0.03
SC	0.09	0.36	0.42	0.34
IB	0.05	0.25	0.24	0.26
FC	0.08	0.35	0.32	0.37

$N_{D,spatial} = 1$

Model weight for / Data generated by

	MD	SC	IB	FC
MD	1.00	0.00	0.00	0.00
SC	0.00	0.63	0.16	0.14
IB	0.00	0.14	0.37	0.34
FC	0.00	0.23	0.47	0.52

$N_{D,spatial} = 3$

Model weight for / Data generated by

	MD	SC	IB	FC
MD	1.00	0.00	0.00	0.00
SC	0.00	1.00	0.00	0.00
IB	0.00	0.00	0.52	0.34
FC	0.00	0.00	0.48	0.66

$N_{D,spatial} = 5$

Fig. 6 Model confusion matrices for calcite content [%] and calcium concentration [mol/m^3] of the three models and the measurement data (MD) over increasing amount of used spatial data points $N_{D,\text{spatial}}$

initial biofilm model M_{IB} and the *full complexity* model M_{FC} confuse their predictions and are not confident in identifying their own predictions (the *initial biofilm* model M_{IB} for calcite is not even able to identify itself). However, only for the *simple chemistry* model M_{SC} the weight is 1.00 and therefore its "level of detail" is perfectly supported with the full data set. The measurement data (MD) obtain a model weight of 1.00 for the full data set too, since it is clearly able to identify itself with the full data set. The weights for the models with the measurement data as the data-generating process are strikingly low. In statistical terms, this means that all models are clearly rejected by the full data set. This fits with the conclusions drawn in [25], that there is at least one relevant process not yet implemented in "sufficient detail", which is necessary for better results.

4.2.2 How much data do we need?

The matrices on the left in Fig. 6 show that considering only one spatial data point is not sufficient, since the diagonal entries for calcite and calcium are all less than 0.50 except for the measurement data for the calcium concentration. This means that there is no "absolute majority" in favor of justifiability for any model and even the measurement data of the calcite content are not able to identify itself (which is obvious since there is clearly a variance between the measurements at different spatial data points). The matrices also show that the simplest model M_{SC} obtains the highest weight of all three models when the data set size is small (principle of parsimony).

When using half of the data set, the simplest model M_{SC} and the most complex model M_{FC} for calcium receive an absolute majority with model weights of 0.63 and 0.52, while the data set size does not suffice for self-identification of the *initial biofilm* model M_{IB}. The weight of M_{IB} on the diagonal entry increases with an increasing data set size, but it never gains a weight greater than 0.5. In contrast, the weight for M_{IB} for the calcium concentration reaches the absolute majority, which means that the data set size is sufficient for self-identification and the physical model assumptions leading to simplifications are justifiable.

Let us now have a closer look on the main-diagonal entries of the model confusion matrix ("self-identification weights") over an increasing data set size in Fig. 7.

It shows, that for the simplest model M_{SC} and clearly for the measurement data, perfect justification (model weight of 1.00) is achieved very quickly. For the *initial biofilm* model M_{IB} and the *full complexity* model M_{FC}, a larger data set size is required to justify their complexity. Since the weights for the more complex models do not stagnate at some point,

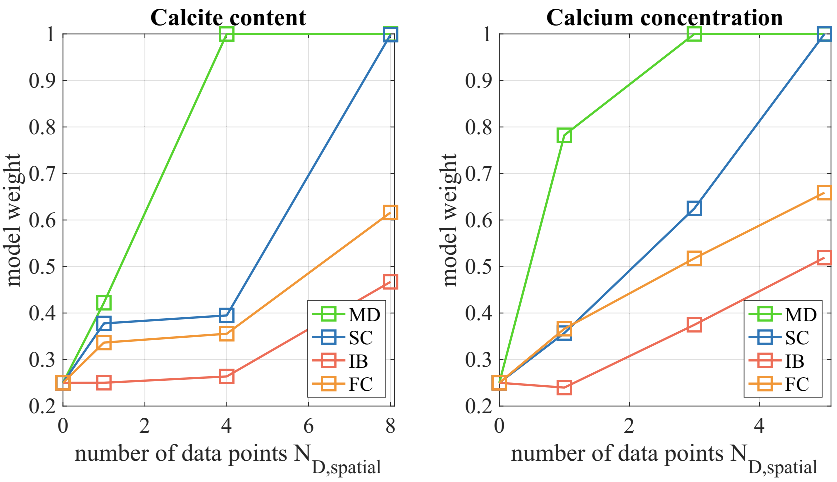


Fig. 7 Average model weights for the data-generating process of the two quantities of interest (calcite content and calcium concentration) of the three models and the measurement data (MD) over increasing amount of used spatial data points $N_{D,\mathrm{spatial}}$

we do not expect that a much larger data set is required to justify their complexity.

When comparing both quantities of interest for the same data set size, the data-generating process for the calcite content is always identified with less confidence (i.e. obtains a lower weight) than for calcium.

4.2.3 How similar are the models?

Now we will assess the similarities between the different models looking on the off-diagonal entries in Fig. 6. For a single data point, we can clearly see that the models “confuse” their predictions, as the off-diagonal weights are relatively high. When the *initial biofilm* model M_{IB} or the *full complexity* model M_{FC} are the data-generating process for the calcite content, the weights for the other models are even larger than the main-diagonal entry. For increasing data set size, the dissimilarities between the models become more significant, but only for the calcium concentration. In contrast, the model confusion remains for the calcium predictions, i.e. the current data set size does not yield a clearer distinction between the models. However, using the full data set, the model confusion decreases significantly, only the similarity between the *initial biofilm* model M_{IB} and the *full complexity* model M_{FC} remains clearly visible. For both calcite and calcium, M_{IB} and M_{FC} are similar, since they both have a relatively high weight, when the other one generated the data. Having a look only at the calcite content shows that even when the *initial biofilm* model M_{IB} is the data-generating process, the *full complexity* model M_{FC} obtains a higher weight, which means that the model cannot be justified with this data set size [68].

4.2.4 How well do the models fit the data?

In a last step, we will analyze the goodness-of-fit of the models to the measurement data. Figure 8 shows the determination coefficient (R^2) between the different model outputs and the measurement data, averaged over all model outputs evaluated on P different collocation points:

$$R^2 = \frac{1}{P}\sum_{i=1}^{P}\left(\frac{\sum_{j=1}^{N_s}\left(\boldsymbol{y}_{0,j}-\overline{\boldsymbol{y}}_0\right)^2}{\sum_{j=1}^{N_s}\left(M_{k,j}\left(\boldsymbol{\omega}^{(i)}\right)-\overline{\boldsymbol{y}}_0\right)^2}\right), \tag{25}$$

with $\boldsymbol{y}_{0,j}$ being the vector of measurements at position j of total length N_s, its mean $\overline{\boldsymbol{y}}_0$ and $M_{k,j}\left(\boldsymbol{\omega}^{(i)}\right)$ the model output of model M_k at position j evaluated at collocation point $\boldsymbol{\omega}^{(i)}$. The R^2 values for different predictions of the same model (different evaluations on different collocation points) were averaged to obtain one representative value per model. For both, calcite content and calcium concentration predictions, the mean R^2 is highest for the *simple chemistry* model M_{SC}. With regard to the BMS analysis it shows that the small BMS weights of the *initial biofilm* model M_{IB} and the *full complexity* model M_{FC} stem from a lower goodness-of-fit and a higher complexity than the *simple chemistry* model M_{SC}. Remember that a more complex model needs a significantly better goodness-of-fit to justify its complexity [68] (and to achieve a similar weight as a simpler model). Furthermore, it is interesting that the weight of the *initial biofilm* model M_{IB} is smaller than the one for the *full complexity* model M_{FC} for the same data set size, although the *full complexity* model M_{FC} is slightly more complex while their goodness-of-fit is similar. Therefore, the high computational effort of the *initial biofilm* model M_{IB} is not justified.

4.2.5 Results

Combining the insights from the Bayesian model justifiability analysis and the goodness-of-fit analysis, we draw the following conclusions about the *initial biofilm* model M_{IB}

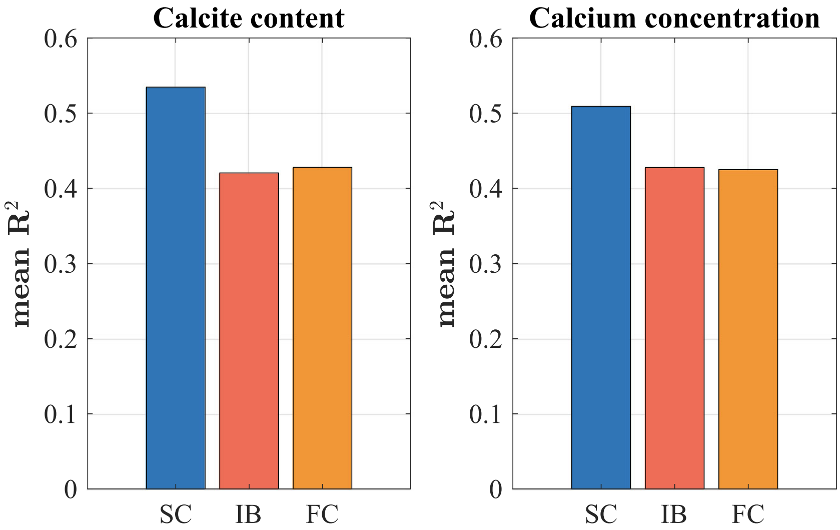


Fig. 8 Mean R^2 between the different model outputs and the measurement data

and *simple chemistry* model M_{SC} as simplifications of the *full complexity* model M_{FC}: The *initial biofilm* model M_{IB} achieves moderate BME values in the BMS analysis and does not use its full potential according to the Bayesian model justifiability analysis. Additionally, M_{IB} provides unsatisfactory goodness-of-fit to the measurement data and cannot capture the underlying physical process reasonably well. The *simple chemistry* model M_{SC} for calcite and calcium obtains the same weight of 1.00 in the BMS analysis (Fig. 7) and Bayesian model justifiability for (Fig. 6) with the full data set. Therefore, the *simple chemistry* model M_{SC} uses its full potential to represent the data and it captures the response of the underlying physical system appropriately.

5 Summary and conclusions

Bayesian model selection (BMS) cannot only be used for ranking models based on their goodness-of-fit to measurement data and parsimony, but also to quantify similarities among models. This work introduces the surrogate-based Bayesian model justifiability analysis for analyzing microbially induced calcite precipitation models in porous media. The suggested framework offers a rigorous pathway to address so-called conceptual uncertainty, i.e. which model is best suited for describing the underlying physical system. The justifiability analysis compares the models among each other and the available measurement data.

Applying the justifiability analysis in addition to the BMS analysis yields a better insight on why a model wins the BMS ranking: either because it fits the measurement data best or only because the data set size is too small to identify a more complex model, that actually fits better. In the latter case, the apparently best model is only best given a too small data set size [68].

The BMS and justifiability analysis were performed using surrogate models, which were built via an arbitrary polynomial chaos expansion (aPC) in order to assure feasibility of the analyses for computationally demanding biogeochemical models. The aPC accelerates the analysis, which requires a large number of model evaluations, by reducing the required number of evaluations of the original model. We apply Bayesian iterative updating of the surrogate models improving their accuracy while incorporating measurement data. In order to account for the error that arises by comparing the surrogates instead of the original models, correction factors for the calculated weights were introduced. The correction factor proposed by [43], correcting the comparison of a model and measurements, was extended to a novel correction factor for a comparison between two computationally demanding models. It helps to perform a reliable surrogate-based Bayesian model justifiability analysis.

Applying the introduced Bayesian model justifiability analysis to three different models (*simple chemistry* model M_{SC},*initial biofilm* model M_{IB} and *full complexity* model M_{FC}), we compare the models to measurement data and among each other. The comparison is based on the predictions of calcite content and calcium concentration at different data points in space and time. The justifiability analysis has shown that the *simple chemistry* model M_{SC} and the *full complexity* model M_{FC} for calcite and calcium and the *initial biofilm* model M_{IB} only for calcium identify themselves best, compared to the other models, when a certain data set size is used. The *simple chemistry* model M_{SC} even achieves perfect justification with a weight of 1.00.

The analysis has also revealed that the data set size is too small for justification of the *initial biofilm* model M_{IB} in terms of the calcium concentration, since its diagonal entries of the model confusion matrix are always smaller than 0.5. Further, it shows that the *initial biofilm* model M_{IB} and the *full complexity* model M_{FC} are similar in terms of both quantities of interest (calcite content and calcium concentration). Additionally, performing the conventional BMS analysis reveals the *simple chemistry* model M_{SC} as the best model in the model set, because of its best trade-off between goodness-of-fit to the measurement data and its sufficiently small degree of complexity.

The proposed analysis provides an extension of the very general justifiability analysis by [68] that makes it applicable for computationally expensive models. It can be concluded that the results for surrogate models followed the intuitively assumed preference for the simplest model when only limited amount of data is available. This makes the method ideal for application cases where the same situation, limited amount of measurement data and computationally expensive models, appears. Although this method poses an effective way of comparing computationally expensive models their computational cost must not be disregarded. With increasing computational cost the number of model evaluations decreases for a given period of time, which leads to a more imprecise surrogate model and therefore less reliable results in the justifiability analysis.

Appendix A: Computational details for the overdetermined system of equations

The solution of the overdetermined system needs to be approximated by minimizing the Euclidian norm (L_2-norm) of the residual:

$$\min_{\mathbf{c}(\mathbf{x},t)} \|\boldsymbol{\Psi}(\boldsymbol{\omega}) \cdot \mathbf{c}(\mathbf{x}, t) - \boldsymbol{M}(\mathbf{x}, t; \boldsymbol{\omega})\|_2.$$

via a linear regression:

$$\boldsymbol{\Psi}^T(\boldsymbol{\omega}) \cdot \boldsymbol{\Psi}(\boldsymbol{\omega}) \cdot \mathbf{c}(\mathbf{x}, t) = \boldsymbol{\Psi}^T(\boldsymbol{\omega}) \cdot \boldsymbol{M}(\mathbf{x}, t; \boldsymbol{\omega}).$$

The new system is determined again and can be solved with the help of the pseudoinverse:

$$\mathbf{c}(\mathbf{x}, t) = \left(\boldsymbol{\Psi}^T(\boldsymbol{\omega}) \cdot \boldsymbol{\Psi}(\boldsymbol{\omega})\right)^{-1} \cdot \boldsymbol{\Psi}^T(\boldsymbol{\omega}) \cdot \boldsymbol{M}(\mathbf{x}, t; \boldsymbol{\omega})$$
$$\mathbf{c}(\mathbf{x}, t) = \boldsymbol{\Psi}^+(\boldsymbol{\omega}) \cdot \boldsymbol{\Psi}^T(\boldsymbol{\omega}) \cdot \boldsymbol{M}(\mathbf{x}, t; \boldsymbol{\omega}),$$

where $\boldsymbol{\Psi}^+(\boldsymbol{\omega})$ denotes the pseudoinverse.

Acknowledgements The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Collaborative Research Center 1253 CAMPOS (DFG, Grant Agreement SFB 1253/1 2017), Collaborative Research Center 1313 (SFB1313) (DFG, Project Number 327154368), DFG project number 380443677 and the Cluster of Excellence EXC 2075 “Data-integrated Simulation Science (SimTech)” at the University of Stuttgart under Germany’s Excellence Strategy - EXC 2075 - 390740016.

Measurement data are available in [25], data for the MICP models and the Bayesian model justifiability analysis is available online in the repository https://git.iws.uni-stuttgart.de/dumux-pub/scheurer2019a.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Alpaydin, E.: Introduction to Machine Learning. Adaptive computation and machine learning. MIT Press, Massachusetts (2004)
2. Baartman, J.E., Melsen, L.A., Moore, D., van der Ploeg, M.J.: On the complexity of model complexity: Viewpoints across the geosciences. CATENA **186**, 10426 (2020). https://doi.org/10.1016/j.catena.2019.104261. https://www.sciencedirect.com/science/article/pii/S0341816219304035
3. Babu, G.J.: Resampling methods for model fitting and model selection. J. Biopharm. Stat. **21**(6), 1177–1186 (2011). https://doi.org/10.1080/10543406.2011.607749
4. Bachmann, R.T., Johnson, A.C., Edyvean, R.G.: Biotechnology in the petroleum industry: an overview. Int. Biodeteriorat. Biodegrad. **86**, 225–237 (2014)
5. Barkouki, T., Martinez, B., Mortensen, B., Weathers, T., De Jong, J., Ginn, T., Spycher, N., Smith, R., Fujita, Y.: Forward and Inverse bio-Geochemical Modeling of Microbially Induced Calcite Precipitation in half-Meter Column Experiments. Transp. Porous Media **90**(1), 23 (2011)
6. Beckers, F., Heredia, A., Noack, M., Nowak, W., Wieprecht, S., Oladyshkin, S.: Bayesian Calibration and Validation of a Large-Scale and Time-Demanding Sediment Transport Model. Water Resourc. Res. **56**(7), e2019WR026966 (2020)
7. Blatman, G., Sudret, B.: An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. Probab. Eng. Mechan. **25**(2), 183–197 (2010)
8. Bottero, S., Storck, T., Heimovaara, T.J., van Loosdrecht, M.C., Enzien, M.V., Picioreanu, C.: Biofilm development and the dynamics of preferential flow paths in porous media. Biofouling **29**(9), 1069–1086 (2013)
9. Brunetti, G., Šimůringnek J, glöckler, D., Stumpp, C.: Handling model complexity with parsimony: Numerical analysis of the nitrogen turnover in a controlled aquifer model setup. J. Hydrol. **584**, 124681 (2020)
10. Burnham, K.P., Anderson, D.R. A Practical Information-Theoretic Approach. Model Selection and Multimodel Inference, 2nd edn. Springer, New York (2002)
11. Cremers, K.J.M.: Stock return predictability: a bayesian model selection perspective. Rev. Financ. Stud. **15**(4), 27 (2002)
12. Cunningham, A.B., Class, H., Ebigbo, A., Gerlach, R., Phillips, A.J., Hommel, J.: Field-scale modeling of microbially induced calcite precipitation. Comput. Geosci. **23**(2), 399–414 (2019)

13. Cuthbert, M.O., McMillan, L.A., Handley-Sidhu, S., Riley, M.S., Tobler, D.J., Phoenix, V.R.: A field and modeling study of fractured rock permeability reduction using microbially induced calcite precipitation. Environ. Sci. Technol. **47**(23), 13637–13643 (2013). https://doi.org/10.1021/es402601g
14. Dupraz, S., Parmentier, M., Ménez, B., Guyot, F.: Experimental and numerical modeling of bacterially induced pH increase and calcite precipitation in saline aquifers. Chem. Geol. **265**(1-2), 44–53 (2009). https://doi.org/10.1016/j.chemgeo.2009.05.003
15. Ebigbo, A., Phillips, A.J., Gerlach, R., Helmig, R., Cunningham, A.B., Class, H., Spangler, L.H.: Darcy-scale modeling of microbially induced carbonate mineral precipitation in sand columns. Water Resour. Res. **48**(7), W07519 (2012). https://doi.org/10.1029/2011WR011714
16. Enemark, T., Peeters, L.J., Mallants, D., Batelaan, O.: Hydrogeological conceptual model building and testing: a review. J. Hydrol. **569**, 310–329 (2019)
17. Gomez, M.G., Anderson, C.M., Graddy, C.M.R., DeJong, J.T., Nelson, D.C., Ginn, T.R.: Large-Scale comparison of bioaugmentation and biostimulation approaches for biocementation of sands. J. Geotechnical Geoenviron. Eng. **143**(5), 04016124 (2017). https://doi.org/10.1061/(ASCE)GT.1943-5606.0001640
18. Gupta, H.V., Clark, M.P., Vrugt, J.A., Abramowitz, G., Ye, M.: Towards a comprehensive assessment of model structural adequacy. Water Resour. Res. 48(8). https://doi.org/10.1029/2011WR011044 (2012)
19. Hamdan, N., Kavazanjian, E. Jr., Rittmann, B.E.: Sequestration of radionuclides and metal contaminants through microbially-induced carbonate precipitation. In: Proc. 14Th Pan American Conf. Soil Mech. Geotech., Engng., Toronto (2011)
20. Head, I.M.: Bioremediation: towards a credible technology. Microbiology **144**(3), 599–608 (1998)
21. Helmig, R.: Multiphase Flow and Transport Processes in the Subsurface - A Contribution to the Modeling of Hydrosystems. Springer, Berlin (1997)
22. Højberg, A., Refsgaard, J.: Model uncertainty – parameter uncertainty versus conceptual models. Water Sci. Technol. **52**(6), 177–186 (2005). https://doi.org/10.2166/wst.2005.0166
23. Höge, M., Wöhling, T., Nowak, W.: A primer for model selection: The decisive role of model complexity. Water Resour. Res. **54**(3), 1688–1715 (2018)
24. Höge, M., Guthke, A., Nowak, W.: The hydrologist's guide to Bayesian model selection, averaging and combination. J. Hydrol. **572**, 96–107 (2019)
25. Hommel, J., Lauchnor, E., Phillips, A., Gerlach, R., Cunningham, A.B., Helmig, R., Ebigbo, A., Class, H.: A revised model for microbially induced calcite precipitation: Improvements and new insights based on recent experiments. Water Resour. Res. **51**(5), 3695–3715 (2015)
26. Hommel, J., Ebigbo, A., Gerlach, R., Cunningham, A.B., Helmig, R., Class, H.: Finding a balance between accuracy and effort for modeling biomineralization. Energy Procedia **97**, 379–386 (2016a)
27. Hommel, J., Lauchnor, E.G., Gerlach, R., Cunningham, A.B., Ebigbo, A., Helmig, R., Class, H.: Investigating the influence of the initial biomass distribution and injection strategies on Biofilm-Mediated calcite precipitation in porous media. Transp. Porous Media **114**(2), 557–579 (2016b). https://doi.org/10.1007/s11242-015-0617-3
28. Hooten, M.B., Hobbs, N.T.: A guide to Bayesian model selection for ecologists. Ecol. Monogr. **85**(1), 3–28 (2015). https://doi.org/10.1890/14-0661.1
29. Huang, S., Cao, M., Cheng, L.: Experimental study on the mechanism of enhanced oil recovery by multi-thermal fluid in offshore heavy oil. Int. J. Heat Mass Transf. **122**, 1074–1084 (2018)
30. Hunter, K.S., Wang, Y., Van Cappellen, P.: Kinetic modeling of microbially-driven redox chemistry of subsurface environments: coupling transport, microbial metabolism and geochemistry. J. Hydrol. **209**(1-4), 53–80 (1998)
31. Jefferys, W.H., Berger, J.O.: Ockham's razor and bayesian analysis. Am. Sci. **80**(1), 64–72 (1992)
32. Kass, R.E., Raftery, A.E.: Bayes factors. J. Amer. Stat. Assoc. **90**(430), 773–795 (1995). https://doi.org/10.1080/01621459.1995.10476572
33. Kirkland, C.M., Thane, A., Hiebert, R., Hyatt, R., Kirksey, J., Cunningham, A.B., Gerlach, R., Spangler, L., Phillips, A.J.: Addressing wellbore integrity and thief zone permeability using microbially-induced calcium carbonate precipitation (MICP): a field demonstration. J. Pet. Sci. Eng. **190**, 107060 (2020). https://doi.org/10.1016/j.petrol.2020.107060
34. Köpel, M., Franzelin, F., Kröker, I., Oladyshkin, S., Santin, G., Wittwar, D., Barth, A., Haasdonk, B., Nowak, W., Pflüger, D., Rohde, C.: Comparison of data-driven uncertainty quantification methods for a carbon dioxide storage benchmark scenario. Comput. Geosci. **23**(2), 339–354 (2019). https://doi.org/10.1007/s10596-018-9785-x
35. Landa-Marbán, D., Tveit, S., Kumar, K., Gasda, S.E.: Practical approaches to study microbially induced calcite precipitation at the field scale. arXiv:201104744 (2020)
36. Lever, J., Krzywinski, M., Altman, N.: Model selection and overfitting. Nat. Methods **13**(9), 703–704 (2016). https://doi.org/10.1038/nmeth.3968
37. Lovley, D.R., Chapelle, F.H.: Deep subsurface microbial processes. Rev. Geophys. **33**(3), 365–381 (1995)
38. MacQuarrie, K.T.B., Mayer, K.U.: Reactive transport modeling in fractured rock: a state-of-the-science review. Earth Sci. Rev. **72**(3-4), 189–227 (2005). https://doi.org/10.1016/j.earscirev.2005.07.003
39. McInerney, M.J., Nagle, D.P., Knapp, R.M.: Microbially enhanced oil recovery: past, Present, and Future. Petroleum Microbiology 215–237 (2005)
40. Megharaj, M., Ramakrishnan, B., Venkateswarlu, K., Sethunathan, N., Naidu, R.: Bioremediation approaches for organic pollutants: a critical perspective. Environ. Int. **37**(8), 1362–1375 (2011)
41. Minto, J.M., Lunn, R.J., El Mountassir, G.: Development of a reactive transport model for Field-Scale simulation of microbially induced carbonate precipitation. Water Resour. Res. **55**(8), 7229–7245 (2019). https://doi.org/10.1029/2019WR025153
42. Mitchell, A.C., Phillips, A.J., Schultz, L., Parks, S., Spangler, L.H., Cunningham, A.B., Gerlach, R.: Microbial $CaCO_3$ mineral formation and stability in an experimentally simulated high pressure saline aquifer with supercritical CO_2. International Journal of Greenhouse Gas Control **15**, 86–96 (2013). https://doi.org/10.1016/j.ijggc.2013.02.001
43. Mohammadi, F., Kopmann, R., Guthke, A., Oladyshkin, S., Nowak, W.: Bayesian selection of hydro-morphodynamic models under computational time constraints. Adv. Water Resour. **117**, 53–64 (2018)
44. Mujah, D., Shahin, M.A., Cheng, L.: State-of-the-art Review of Biocementation by Microbially Induced Calcite Precipitation (MICP) for Soil Stabilization. Geomicrobiol. J. **34**(6), 524–537 (2017). https://doi.org/10.1080/01490451.2016.1225866
45. Mulligan, C.N., Galvez-Cloutier, R.: Bioremediation of metal contamination. Environ. Monit. Assess. **84**(1-2), 45–60 (2003)
46. Nassar, M.K., Gurung, D., Bastani, M., Ginn, T.R., Shafei, B., Gomez, M.G., Graddy, C.M., Nelson, D.C., DeJong, J.T.: Large-Scale Experiments in Microbially Induced Calcite Precipitation (MICP): Reactive Transport Model Development and Prediction. Water Resour. Res. **54**(1), 480–500 (2018)

47. Nearing, G.S., Gupta, H.V.: Ensembles vs. information theory: supporting science under uncertainty. Frontiers of Earth Science **12**(4), 653–660 (2018)
48. Neuman, S.P.: Maximum likelihood bayesian averaging of uncertain model predictions. Stoch. Env. Res. Risk A. **17**(5), 291–305 (2003)
49. Oladyshkin, S.: aPC Matlab Toolbox: Data-driven Arbitrary Polynomial Chaos, Matlab Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/72014-apc-matlab-toolbox-data-driven-arbitrary-polynomial-chaos (2020a)
50. Oladyshkin, S.: BaPC Matlab Toolbox: Bayesian Arbitrary Polynomial Chaos, Matlab Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/74006-bapc-matlab-toolbox-bayesian-arbitrary-polynomial-chaos (2020b)
51. Oladyshkin, S., Nowak, W.: Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. Reliab. Eng. Syst. Safe. **106**, 179–190 (2012)
52. Oladyshkin, S., de Barros, F., Nowak, W.: Global sensitivity analysis: a flexible and efficient framework with an example from stochastic hydrogeology. Adv. Water Resour. **37**, 10–22 (2012)
53. Oladyshkin, S., Class, H., Nowak, W.: Bayesian updating via bootstrap filtering combined with data-driven polynomial chaos expansions: methodology and application to history matching for carbon dioxide storage in geological formations. Comput. Geosci. **17**(4), 671–687 (2013a)
54. Oladyshkin, S., Schröder, P., Class, H., Nowak, W.: Chaos Expansion based Bootstrap Filter to Calibrate CO_2 Injection Models. Energy Procedia **40**, 398–407 (2013b)
55. Oladyshkin, S., Mohammadi, F., Kroeker, I., Nowak, W.: Bayesian^3 active learning for the gaussian process emulator using information theory. Entropy **22**(8), 890 (2020)
56. van Paassen, L.A., Ghose, R., van der Linden, T.J.M., van der Star, W.R.L., van Loosdrecht, M.C.M.: Quantifying Biomediated Ground Improvement by Ureolysis: Large-Scale Biogrout Experiment. J. Geotechnical Geoenviron. Eng. **136**(12), 1721–1728 (2010). https://doi.org/10.1061/(ASCE)GT.1943-5606.0000382
57. Parkinson, D., Mukherjee, P., Liddle, A.R.: Bayesian model selection analysis of wMAP3. Phys Rev D **73**, 123523 (2006). https://doi.org/10.1103/PhysRevD.73.123523
58. Phillips, A.J., Lauchnor, E., Eldring, J., Esposito, R., Mitchell, A.C., Gerlach, R., Cunningham, A.B., Spangler, L.H.: Potential CO_2 Leakage Reduction Through Biofilm-induced calcium carbonate precipitation. Environ. Sci. Technol. **47**(1), 142–149 (2013)
59. Phillips, A.J., Cunningham, A.B., Gerlach, R., Hiebert, R., Hwang, C., Lomans, B.P., Westrich, J., Mantilla, C., Kirksey, J., Esposito, R., Spangler, L.H.: Fracture sealing with Microbially-Induced calcium carbonate precipitation: a field study. Environ. Sci. Technol. **50**, 4111–4117 (2016). https://doi.org/10.1021/acs.est.5b05559
60. Raftery, A.E.: Bayesian model selection in social research. Sociol. Methodol. 111–163 (1995)
61. Refsgaard, J.C., Christensen, S., Sonnenborg, T.O., Seifert, D., Højberg, A.L., Troldborg, L.: Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. Adv. Water Resour. **36**, 36–50 (2012)
62. Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resour. Res. 46(5). https://doi.org/10.1029/2009WR008328 (2010)
63. Rojas, R., Feyen, L., Dassargues, A.: Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. Water Resour. Res. 44(12). https://doi.org/10.1029/2008WR006908 (2008)
64. Rojas, R., Kahunde, S., Peeters, L., Batelaan, O., Feyen, L., Dassargues, A.: Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling. J. Hydrol. **394**(3-4), 416–435 (2010)
65. Schäfer Rodrigues Silva, A., Guthke, A., Höge, M., Cirpka, O.A., Nowak, W.: Strategies for simplifying reactive transport models - a Bayesian model comparison. Water Res. Res. p e2020WR028100. https://doi.org/10.1029/2020WR028100 (2020)
66. Schmidt, E.: Zur theorie der linearen und nichtlinearen integralgleichungen. In: Integralgleichungen und Gleichungen mit unendlich vielen Unbekannten, pp. 190–233. Springer (1989)
67. 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**(12), 9484–9513 (2014)
68. Schöniger, A., Illman, W., Wöhling, T., Nowak, W.: Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. J. Hydrol. **531**, 96–110 (2015a)
69. Schöniger, A., Wöhling, T., Nowak, W.: A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. Water Resour. Res. **51**(9), 7524–7546 (2015b)
70. Steefel, C., MacQuarrie, K.: Reactive transport in porous media. Reviews in mineralogy, mineralogical society of america. Washington, chap Approaches to modelling of reactive transport in porous media 82–129 (1996)
71. Steefel, C., Depaolo, D., Lichtner, P.: Reactive transport modeling: An essential tool and a new research approach for the Earth sciences. Earth Planet. Sci. Lett. **240**(3-4), 539–558 (2005). https://doi.org/10.1016/j.epsl.2005.09.017
72. Stocks-Fischer, S., Galinat, J.K., Bang, S.S.: Microbiological precipitation of $CaCO_3$. Soil Biol. Biochem. **31**, 1563–1571 (1999). https://doi.org/10.1016/S0038-0717(99)00082-6
73. Suliman, F., French, H., Haugen, L., Søvik, A.: Change in flow and transport patterns in horizontal subsurface flow constructed wetlands as a result of biological growth. Ecologic. Eng. **27**(2), 124–133 (2006)
74. Terzis, D., Laloui, L.: A decade of progress and turning points in the understanding of bio-improved soils: a review. Geomechan. Energ. Environ. **19**, 100116 (2019)
75. Troldborg, L., Refsgaard, J.C., Jensen, K.H., Engesgaard, P.: The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. Hydrogeol. J. **15**(5), 843–860 (2007)
76. Umar, M., Kassim, K.A., Chiet, K.T.P.: Biological process of soil improvement in civil engineering: A review. J. Rock Mechan. Geotechnic. Eng. **8**(5), 767–774 (2016). https://doi.org/10.1016/j.jrmge.2016.02.004. http://www.sciencedirect.com/science/article/pii/S1674775516300245
77. Villadsen, J., Michelsen, M.: Solution of Differential Equation Models by Polynomial Approximation, vol. 7. Prentice-Hall, Englewood Cliffs (1978)
78. van der Vorst, H.A.: BI-CGSTAB: A fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **13**(2), 631–644 (1992). https://doi.org/10.1137/0913035
79. Wasserman, L.: Bayesian model selection and model averaging. J. Math. Psychol. **44**(1), 92–107 (2000)
80. Whiffin, V.S., La, v.an.P.ssen, Harkes, M.P.: Microbial carbonate precipitation as a soil improvement technique. Geomicrobiol J. **24**(5), 417–423 (2007). https://doi.org/10.1080/01490450701436505
81. Wöhling, T., Schöniger, A., Gayler, S., Nowak, W.: Bayesian model averaging to explore the worth of data for soil-plant model

selection and prediction. Water Resour. Res. **51**(4), 2825–2846 (2015). https://doi.org/10.1002/2014WR016292

82. Wiener, N.: The homogeneous chaos. Am. J. Math. **60**(4), 897–936 (1938). https://doi.org/10.2307/2371268
83. van Wijngaarden, W.K., van Paassen, L.A., Vermolen, F.J., van Meurs, G.A.M., Vuik, C.: A reactive transport model for biogrout compared to experimental data. Transp. Porous Media **111**(3), 627–648 (2016). https://doi.org/10.1007/s11242-015-0615-5
84. Xiu, D., Karniadakis, G.E.: Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. Comput. Methods Appl. Mechan. Eng. **191**(43), 4927–4948 (2002a)
85. Xiu, D., Karniadakis, G.E.: The wiener–askey polynomial chaos for stochastic differential equations. SIAM J. Scientif. Comput. **24**(2), 619–644 (2002b)
86. Xu, T., Sonnenthal, E., Spycher, N., Pruess, K.: TOUGHREACT - A simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media: Applications to geothermal injectivity and CO_2 geological sequestration. Comput. Geosci. **32**(2), 145–165 (2006). https://doi.org/10.1016/j.cageo.2005.06.014
87. Yang, Y., Chu, J., Cao, B., Liu, H., Cheng, L.: Biocementation of soil using non-sterile enriched urease-producing bacteria from activated sludge. J. Clean. Prod. **262**, 121315 (2020). https://doi.org/10.1016/j.jclepro.2020.121315

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Stefania Scheurer[1] · **Aline Schäfer Rodrigues Silva**[1] · **Farid Mohammadi**[2] · **Johannes Hommel**[2] · **Sergey Oladyshkin**[1] · **Bernd Flemisch**[2] · **Wolfgang Nowak**[1]

[1] Department of Stochastic Simulation and Safety Research for Hydrosystems (IWS/SimTech), University of Stuttgart, 70569 Stuttgart, Germany

[2] Department of Hydromechanics and Modelling of Hydrosystems (IWS), University of Stuttgart, 70569 Stuttgart, Germany