Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR028100

Key Points:

- We compare a set of four simplified models against a reference model for reactive transport at quasi steady state on aquifer scale
- A Bayesian model justifiability analysis helps identifying the most suitable model simplification strategy
- The proposed analysis reveals the difficulty of reasonably constraining parameter priors for simplified models

Correspondence to:
A. Schäfer Rodrigues Silva,
aline.schaefer@iws.uni-stuttgart.de

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

Received 10 JUN 2020
Accepted 13 OCT 2020
Accepted article online 24 OCT 2020

Strategies for Simplifying Reactive Transport Models: A Bayesian Model Comparison

Aline Schäfer Rodrigues Silva[1], Anneli Guthke[1], Marvin Höge[1], Olaf A. Cirpka[2], and Wolfgang Nowak[1]

[1]Department of Stochastic Simulation and Safety Research for Hydrosystems (IWS/SimTech),University of Stuttgart,Stuttgart,Germany , [2]Center for Applied Geoscience,University of Tübingen,Germany

Abstract For simulating reactive transport on aquifer scale, various modeling approaches have been proposed. They vary considerably in their computational demands and in the amount of data needed for their calibration. Typically, the more complex a model is, the more data are required to sufficiently constrain its parameters. In this study, we assess a set of five models that simulate aerobic respiration and denitrification in a heterogeneous aquifer at quasi steady state. In a probabilistic framework, we test whether simplified approaches can be used as alternatives to the most detailed model. The simplifications are achieved by neglecting processes such as dispersion or biomass dynamics, or by replacing spatial discretization with travel-time-based coordinates. We use the model justifiability analysis proposed by Schöniger, Illman, et al. (2015, https://doi.org/10.1016/j.jhydrol.2015.07.047) to determine how similar the simplified models are to the reference model. This analysis rests on the principles of Bayesian model selection and performs a tradeoff between goodness-of-fit to reference data and model complexity, which is important for the reliability of predictions. Results show that, in principle, the simplified models are able to reproduce the predictions of the reference model in the considered scenario. Yet, it became evident that it can be challenging to define appropriate ranges for effective parameters of simplified models. This issue can lead to overly wide predictive distributions, which counteract the apparent simplicity of the models. We found that performing the justifiability analysis on the case of model simplification is an objective and comprehensive approach to assess the suitability of candidate models with different levels of detail.

Plain Language Summary In groundwater, chemical substances like nitrate are transported and undergo chemical reactions. Understanding such reactive transport processes plays a key role in securing our water resources and drinking water. We use computer models for understanding such reactive transport processes and for simulating their future behavior. In such models, we make many scientific decisions on which processes should be included and in what degree of detail. Here, we face a trade-off: Usually, a complex model with many mathematical terms resolves many details of the process. Yet, such complex models require lots of data for calibration and lots of time for the computer simulation. In contrast, a simple model with fewer details comes with less effort in both respects. However, it might neglect important parts of the process. For the example of nitrate decay, we use a probabilistic approach to find the best simplification for a comparatively detailed reference model. Our results show that, in certain cases, it is justified to employ a simpler model instead of a complex alternative without deteriorating modeling results. Alongside, we explain how difficult it can be to define realistic parameter ranges for simplified models.

1. Introduction

Our system understanding in environmental science will always remain incomplete, because we can neither fully resolve the spatial and temporal variability of all system properties, nor can we know all processes and their interactions to achieve a description of the true system behavior. This lack of system understanding leads to so-called conceptual uncertainty, which is the uncertainty in choosing the most adequate representation of the real system. Acknowledging that we can only approximate the natural system, we can, however, formulate different models with different degrees or types of simplifications and treat these models as hypotheses about the true system. Though it is impossible to quantify the total conceptual uncertainty

(Höge et al., 2019; Nearing & Gupta, 2018), considering different hypotheses can, at least, help us to estimate how the modeling results may differ depending on our model choice (Ferré, 2017).

Conceptual uncertainty has received increasing attention as it has been identified as a main source of uncertainty in modeling (Burnham & Anderson, 2002; Enemark et al., 2019; Gupta et al., 2012; Neuman, 2003; Refsgaard et al., 2012; Rojas et al., 2008, 2010; Schöniger, Wöhling, et al., 2015; Troldborg et al., 2007). Examples for conceptual uncertainty in reactive transport models are whether mechanisms such as transverse mixing or the growth and decay of biomass are controlling the process on the relevant spatial and temporal scale or whether they can be neglected (Loschko et al., 2016; Sanz-Prat, Lu, Finkel, et al., 2016). Another example for conceptual uncertainty in reactive transport simulation is the choice of a model for the reaction kinetics. This is investigated in a recent study of Brunetti et al. (2020) who compared models for ammonification and nitrification on different levels of complexity.

Apart from the uncertainty about which processes should be included in the conceptual model, data scarcity also restricts computational models in the level of complexity that should be used to describe these processes (e.g., Guthke et al., 2017). Here, it is important to note that the term "model complexity" is not uniquely defined and we refer to Höge et al. (2018) for a detailed discussion of this issue. In a recent study, Baartman et al. (2020) investigated the geoscientific community's understanding of model complexity. Their survey shows that there is "no general consensus on how model complexity is perceived or should be defined." However, 78% of the participants consider the "number of processes explicitly included" as an adequate characterization of model complexity, followed by the "number of interactions/feedback incorporated."

Generally, models with many parameters and nonlinear interactions require more (informative) data to constrain their parameters during calibration. Therefore, model complexity and the number and quality of field data have to be balanced. Typically, if the model is too complex for the given number of calibration data, it will show a good fit during calibration, but a high variance and errors in the predictions beyond calibration conditions. This effect is known as overfitting (Babu, 2011; Lever et al., 2016). Contrarily, a model that is too simple needs less data for calibration but shows a high systematic bias between its predictions and measured data and thus "underfits" the system (Babu, 2011; Lever et al., 2016). This issue is well-known as "bias-variance-tradeoff" (Burnham & Anderson, 2002; Geman et al., 1992). Consequently, for a realistic number of measurements, there is a certain level of model complexity which is just complex enough to capture the variability in the data but not too complex so the model does not overfit ("principle of parsimony") (e.g., Jefferys & Berger, 1992).

Bayesian model selection (BMS) (e.g., Raftery, 1995; Wasserman, 2000) is a statistical method known to yield a model ranking that implicitly reflects an optimal trade-off between model performance and parsimony. This analysis ranks the considered models based on Bayesian Model Evidence (BME), which is an integral measure of how well a model fits a given data set over its entire parameter space (Schöniger et al., 2014).

Here, we use an analysis based on this method to test whether simplified approaches can be used as alternatives to the most detailed model. If we picked a certain set of parameters to run the reference model and used the corresponding predictions as reference data set, we could use BMS to identify the simplified model that achieves the best tradeoff between goodness-of-fit and complexity. But if we slightly changed the reference model's parameter values, our conclusions might change significantly. Given that parameter uncertainty can take up a significant portion of the overall uncertainty in modeling, we need a method that selects the best replacement model in view of the full predictive distribution of the reference model. This can be achieved within the framework of the so-called model justifiability analysis (Schöniger, Illman, et al., 2015).

The core idea of the justifiability analysis is that the models are tested against each other, asking the following: "How would the models be ranked if one of the models actually generated the data?" The original purpose of the method is to identify the justifiable level of complexity given a specific amount and type of data (Schöniger, Illman, et al., 2015). In this study, however, we are interested in how similar the predictions of simplified models are compared to the reference model. For this purpose, we use the justifiability analysis to answer how the simplified models score through the eyes of BMS if the data are generated by the reference model. To represent each model's parameter and prediction space, the method is established in a Monte Carlo framework using random sampling of the parameter prior distributions. We analyze the decisiveness of the resulting model ranking with the so-called Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995). This factor shows whether there is significant evidence for selecting or discarding a model from the set.

Note that our analysis does not test how well models fit real measurements. It should be rather seen as one of two complementary parts of model testing: In the analysis presented here, plausible model alternatives are tested against each other in a synthetic setup to check how different modeling hypotheses affect the prediction of the quantities of interest and which amount and type of data is needed to distinguish between the different models. With this analysis, we aim to eliminate model candidates that are overly complex or simple. For a comprehensive model testing, this analysis can be complemented with a test against measured data. The agreement with actually observed data can be tested with BMS.

While statistical model selection techniques have received growing interest in many disciplines (Cremers, 2002; Hooten & Hobbs, 2015; Raftery, 1995; Schöniger et al., 2014), only Brunetti et al. (2020) have used BMS to identify an appropriate level of complexity for biogeochemical models. Based on transient laboratory measurements, the authors compared five models that differed in the description of the reaction kinetics. To build a model set of varying complexity, they used different combinations of first-order decay laws and Monod kinetics for ammonification and nitrification. They found that Monod kinetics are the best suited choice for modeling this lab-scale experiment. However, they emphasize that the model choice depends on the temporal stage of the experiment: While bacterial growth was a dominating process at the beginning of the experiment (supporting the model with Monod kinetics), it was negligible after a while, thus the process could be described by simpler models with first-order kinetics.

The present study considers reactive transport models of different complexity and assesses in a probabilistic framework how well the different simplified models can mimic the system behavior of a computationally expensive reference model. We investigate a set of five models that simulate aerobic respiration and denitrification in a heterogeneous aquifer at quasi steady state, that is, in a regime for which Brunetti et al. (2020) considered first-order kinetics justifiable. However, in contrast to the latter authors we consider spatial distributions and the interaction between reactive turnover and physical transport.

We use the model justifiability analysis proposed by Schöniger, Illman, et al. (2015) to assess the following research questions: Are the simplified models sufficiently unbiased and flexible enough to reproduce the entire predictive distribution of the computationally expensive reference model? Can we select a simplified model that represents the reference model best or discard a model that performs unsatisfyingly?

With this analysis, we address the specific problem of choosing a model from a fixed set in the presence of parameter uncertainty. We are not concerned with a full uncertainty assessment, considering, for example, the uncertainty in the underlying flow field due to uncertain hydraulic parameters. We also emphasize that the goal is not to quantify conceptual uncertainty, since this is logically impossible for a finite set of model alternatives (e.g., Höge et al., 2019; Nearing & Gupta, 2018).

In summary, our proposed method ranks pre-selected simplified models considering their complete distribution of possible parameter values by identifying the optimal Bayesian tradeoff between performance (agreement with reference data) and parsimony. This systematic assessment of model versions is a novel extension of the justifiability analysis in the context of model simplification. The paper is structured as follows: We present the methods in section 2, starting with the introduction of BMS in subsection 2.1 as the basis for the model justifiability analysis in section 2.2. The different reactive transport models and their underlying assumptions are explained in section 3, followed by details on setup and implementation in section 4. Results are presented and discussed in section 5. We summarize our findings and provide conclusions in section 6.

2. Methods

2.1. BMS

BMS (e.g., Raftery, 1995; Wasserman, 2000) is a well-known approach to address conceptual uncertainty. For this method, it is assumed that the data generating process is contained in the model set (Bernardo et al., 1999; Vehtari & Ojanen, 2012). This assumption is appropriate in our study, as we compare a set of models against a predefined “true” reference model.

In the BMS framework, model weights are calculated that reflect the probability of each model to be the true one. In the limit of infinite data set size, BMS will identify this true model by assigning it a weight of 100% (Höge et al., 2019). In the case of finite data, however, the identification of the true model may be impossible because two or more models receive similar weights (Schöniger, Illman, et al., 2015).

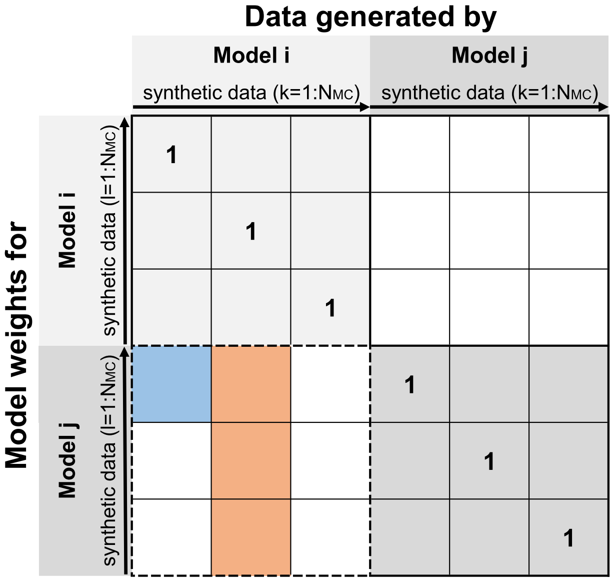


Figure 1. Schematic illustration of the model confusion matrix for two models, *i* and *j*. Blue box: likelihood of a single realization drawn from Model *j* given a realization drawn from Model *i*. Red box: BME value (average likelihood) of Model *j* given a single realization *k* of Model *i*. This BME value is normalized by the sum of the BME values of all models for this data set *k*, which yields a single model weight w_{jk}. Dashed box: Averaging these weights over all synthetic data sets of the generating Model *i* yields the model weight w_j, that is, the expected weight of Model *j* given that Model *i* is true.

As a starting point, prior weights $P(M_i)$ are formulated. In Bayesian statistics, a prior probability reflects the modeler's belief based on expert knowledge. It is formulated before measurements (or synthetic reference data) $\mathbf{y}_0$ is taken into account. In the BMS framework, a typical choice are uniform prior weights $P(M_i) = \frac{1}{N_m}$ that treat all models in the set as equally likely.

After formulating prior weights $P(M_i)$, they are updated to posterior weights $P(M_i|\mathbf{y}_0)$based on Bayes' theorem:

$$P(M_i|\mathbf{y}_0) = \frac{p(\mathbf{y}_0|M_i)\,P(M_i)}{\sum_{j=1}^{N_m} p(\mathbf{y}_0|M_j)\,P(M_j)}, \tag{1}$$

in which $p(\mathbf{y}_0|M_i)$ is the so-called BME. BME is also known as marginal likelihood because it can be calculated by averaging (marginalizing) over the model's parameter space $\mathbf{u}_i$ (Kass & Raftery, 1995; Schöniger et al., 2014):

$$p(\mathbf{y}_0|M_i) = \int_{\mathbf{u}_i} p(\mathbf{y}_0|M_i, \mathbf{u}_i)\,p(\mathbf{u}_i|M_i)\,d\mathbf{u}_i. \tag{2}$$

BME thus quantifies the model's average likelihood to have generated the data $\mathbf{y}_0$ independent of the parameter choice. Equation 2 can be evaluated by sampling the prior distribution of the model parameters $p(\mathbf{u}_i|M_i)$ using N_{MC} Monte Carlo samples and evaluating the likelihood of the reference data $\mathbf{y}_0$ given the predictions based on the parameter vector $\mathbf{u}_i$ of the model M_i (Schöniger et al., 2014):

$$p(\mathbf{y}_0|M_i) \approx \frac{1}{N_{MC}} \sum_{k=1}^{N_{MC}} p(\mathbf{y}_0|M_i, \mathbf{u}_{ik}), \tag{3}$$

where $k = 1, \ldots, N_{MC}$ enumerates the Monte Carlo realizations, so that $\boldsymbol{u}_{ik}$ is parameter realization k for model M_i. This integral over the model's parameter space ensures an optimal tradeoff between goodness-of-fit and parsimony. A narrow, bias-free predictive distribution will obtain a high BME value, while both a very wide distribution and a heavily biased (but narrow) distribution will be punished with a lower value.

2.2. Model Justifiability Analysis

In the model justifiability analysis introduced by Schöniger, Illman, et al. (2015), a set of models are mutually tested against each other by building the so-called "model confusion matrix" (cf. Figure 1). Confusion matrices are often used in machine learning, particularly in the field of statistical classification (e.g., Alpaydin, 2004). It is a special type of contingency table, which compares the actual with the predicted classification. Thus, it is easily visible whether an object is misclassified ("confused").

This concept has been transferred to the problem of model identification: we let the models take turns in generating the data set $\mathbf{y}_0$, which serves as "synthetic truth," and then evaluate how well each data generating model can be identified through the eyes of BMS. If one of the non–data generating models receives a nonnegligible Bayesian model weight (Equation 1), this can be seen as "confusion."

Implementation-wise, we let each of the N_M models generate N_{MC} data sets $\mathbf{y}_0$ by drawing random samples from their prior parameter distributions $p(\mathbf{u}_i|M_i)$. Running the models with these parameters yields predictive distributions. These predictions are treated as synthetic truth $\mathbf{y}_0$. In this role, we refer to the models as "data generating" and list them in the column labels of the model confusion matrix (Figure 1).

Then, all predictions $\mathbf{y}$ of each model (including the one that generated the synthetic truth) are compared to the reference data set $\mathbf{y}_0$. We refer to these models as "evaluating" and list them in the row labels of the matrix. To compare a single data set pair of the "generating" and the "evaluating" model (blue box in Figure 1), we calculate the likelihood $p(\mathbf{y}_{0,ik}|M_j, \mathbf{u}_{jl})$ of the reference data set $\mathbf{y}_{0,\,ik}$ generated by Model M_i with the parameter vector $\mathbf{u}_{ik}$ given the evaluating Model M_j with the parameter vector $\mathbf{u}_{jl}$.

Table 1
Interpretation of Bayes Factors According to Kass and Raftery (1995)

$log_{10}(BF)$	Evidence against M_j
0–0.5	not worth more than a bare mention
0.5–1	substantial
1–2	strong
>2	decisive

Averaging the likelihoods over all parameter realizations of the evaluating model (i.e., averaging over N_{MC} rows) yields a BME value (see Equation 2 and red box in Figure 1). Based on the BME values of all models, we calculate their model weights $P\left(M_j|\mathbf{y}_{0,ik}\right)$ for a single realization k of the data generating model M_i according to Equation 1. These model weights depend on the reference data set $\mathbf{y}_{0,\,ik}$ chosen from the data generating model (column). Therefore, we also average over all N_{MC} columns of the data generating model to obtain the average weight of Model j given the data produced by Model i (dashed box in Figure 1).

The resulting confusion matrix has the size $N_M \times N_M$. Its main-diagonal elements are the so-called self-identification weights and can be used for assessing the justifiability of the models' complexity (see Schöniger, Illman, et al., 2015). The off-diagonal elements can be interpreted as a measure of similarity between two models. For an infinite data set size, the true model will be identified with a weight of 100% and all other models will receive a weight of 0 (Schöniger, Illman, et al., 2015). For finite data sets, we can observe that the models "confuse" their own predictions with the ones of the competing models. In this study, we are interested in the similarity of the simplified models to the complex reference model. Therefore, we focus on the model weights in the first column of the model confusion matrix. We repeat the calculation of model weights over growing data sets. Please note that, in this analysis, the data are not based on field or lab experiments. Thus, the number of data points is only limited by the grid resolution and not by the effort of acquiring field or lab measurements.

2.3. Bayes Factor

Another possibility to analyze BME values is as Bayes Factors (Jeffreys, 1961; Kass & Raftery, 1995), which reflects the decisiveness of model choice. The Bayes factor quantifies the evidence of one model M_i (in our analysis the reference model, denoted as M1) compared to an alternative M_j (in our analysis the four simplified models, denoted as M2–M5).

The Bayes Factor between two models is defined as the ratio of their respective BME values. It can be obtained from posterior odds, that is, the ratio of posterior model weights according to Equation 1, multiplied with the models' prior odds:

$$BF\left(M_i, M_j\right) = \frac{P\left(M_i|\mathbf{y}_0\right)}{P\left(M_j|\mathbf{y}_0\right)} \frac{P\left(M_j\right)}{P\left(M_i\right)} = \frac{P\left(\mathbf{y}_0|M_i\right)}{P\left(\mathbf{y}_0|M_j\right)}. \tag{4}$$

Jeffreys (1961) introduced categories for interpreting the Bayes Factor as evidence against M_j (here: evidence against the simplified models). We will use the slightly modified scale suggested by Kass and Raftery (1995) as shown in Table 1.

Accordingly, negative $log_{10}(BF)$ values favor M_j over M_i (here: the simplified over the reference model).

We calculate Bayes factors for each data set realization generated by the reference model and evaluate the resulting cumulative distribution functions of Bayes factors.

3. Description of the Models

We use the model justifiability analysis to test whether four simplified models are suitable alternatives to the most detailed reference model for simulating aerobic respiration and denitrification in a heterogeneous aquifer. Table 2 gives an overview of the models and in the sections 3.1 to 3.5, we describe the details of each model's conceptualization and their underlying assumptions. Further details of the models can be found in Sanz-Prat et al. (2015), Sanz-Prat, Lu, Amos, et al. (2016), and Loschko et al. (2016).

The considered models are based on different conceptualizations and partly differ considerably in their computational costs. The most complex reference model (M1) is a spatially explicit advection-dispersion-reaction model with biomass growth and decay of a facultative anaerobic organism and transport of dissolved oxygen, nitrate, and dissolved organic carbon (DOC). The DOC is released from the aquifer matrix. From a biogeochemical perspective this is already a highly simplified model as it neglects the reactive intermediates nitrite, nitric oxide, and nitrous oxide, as well as the presence and interactions of different organisms.

Table 2
Overview of the Model Set

Model	Spatial conceptualization	Considered processes	Number of parameters	Run time (s)
$M1$	spatially explicit	dispersion, dynamic biomass	10	38.7
$M2$	spatially explicit	dynamic biomass (Monod)	10	33.3
$M3$	spatially explicit	dispersion	10	31.3
$M4$	streamline based	cum. rel. reactivity (Monod)	5	0.1
$M5$	streamline based	cum. rel. reactivity (zeroth-,first-order decay)	2	0.1

Note. Runtimes are averaged over 10,000 runs on a standard computer with IntelCore i7 CPU @ 3.60 GHz, 32GB RAM.

In contrast to our reference model, many aquifer- or catchment-scale models on nitrate transport neglect almost all details of the reactive system and describe denitrification as a simple first-order decay process that may depend on the organic carbon content of the soil (e.g., Almasri & Kaluarachchi, 2007; Liu et al., 2018; Zhang et al., 2020), even abandoning inhibition of denitrification by dissolved oxygen. The notion of these models is that mechanistic details of the reactions are averaged out in large-scale applications and an effective first-order rate law emerges.

We will test simplified reaction models that stand somewhere between the reference model and first-order laws. As a first approach of simplification, we neglect mixing due to dispersion (M2) and assume that the mixing of electron donors (DOC) and acceptors (dissolved oxygen and nitrate) is mainly caused by mass transfer between the immobile matrix and mobile groundwater. We thereby follow the paradigm of stochastic-convective transport (e.g., Atchley et al., 2013; Dagan & Nguyen, 1989).

We take a second approach to simplification by neglecting biomass growth and decay (M3): as demonstrated by Sanz-Prat et al. (2015), Sanz-Prat, Lu, Amos, et al. (2016) and Loschko et al. (2016b), dynamic biomass growth may not be needed in reactive transport models of dissolved oxygen and nitrate if longer times of nitrate loading are considered. In essence, bacteria grow so fast that abiotic controls, namely, the kinetics of electron donor release from the aquifer matrix, take over. The inhibition by dissolved oxygen, however, suppresses denitrification in young groundwater and should not be neglected.

Under the assumptions discussed for Model M2, self-organization of reactive zones according to advective travel times or times of exposure to reactive aquifer material have been claimed (e.g., Sanz-Prat, Lu, Amos, et al., 2016). This means that, even though biomass, reactive turnover and solute concentrations depend on each other and on physical transport in a seemingly complex way, spatial patterns naturally evolve that are associated with travel or exposure times. Exploiting these conditions leads to our model simplifications M4 and M5: These models are not spatially explicit but are based on the so-called cumulative relative reactivity approach of Loschko et al. (2016). This method replaces the time in the reaction equation with the travel time of a water parcel through the aquifer and accounts for varying reactivity along the travel path. This simplification is only valid if certain assumptions, like a diffusive source of the considered substance, are fulfilled (Loschko et al., 2016). In M4, the aerobic respiration and denitrification are described by standard Monod kinetics with noncompetitive inhibition of denitrification while oxygen is present. M5 uses simplified reaction kinetics compared to M4: Aerobic respiration is described by zeroth-order decay and denitrification is modeled by first-order decay.

In the following section, we describe the details of each model's conceptualization and the assumptions that underlie their simplifications. The values we chose for the fixed parameters are listed in Table A1, the distributions that are used for sampling the parameters that are considered uncertain are given in Table A4 for M1–M3 and in Table A5 for M4 and M5.

3.1. Model M1: Reference Model

Model M1 solves the classical advection-dispersion-reaction equation of the dissolved species i (e.g., Loschko et al., 2016; Steefel & Lichtner, 1998):

$$\frac{\partial c_i}{\partial t} + \mathbf{v} \cdot \nabla c_i - \nabla \cdot \left(\mathbf{D} \nabla c_i \right) = r_i \left(\mathbf{c} \left(\mathbf{x}, t \right), \mathbf{x}, t \right), \tag{5}$$

in which c_i (mol/L) is the concentration of the dissolved species i, which depends on both the spatial coordinates x (m) and time t (s); $\mathbf{v}$ (m/s) denotes the linear average velocity; $\mathbf{D}$ (m^2/s) is the dispersion tensor, and r_i (mol/[ls]) is the reaction rate of component i, which potentially depends on all concentrations, the spatial position and time.

For each immobile component j, the concentration change is given by

$$\frac{\partial c_j^*}{\partial t} = \mathbf{r}^* \left(\mathbf{c} \left(\mathbf{x}, t \right), \mathbf{c}^* \left(\mathbf{x}, t \right) \right). \tag{6}$$

In this study, the dissolved species are oxygen (O_2), nitrate (NO_3^-) and DOC. The concentrations of these mobile species are denoted $c_i\left(\mathbf{x}, t\right)$ (mol/L) whereas the concentration of the immobile species in the soil matrix are given in moles of carbon per volume of water $c_j^*\left(\mathbf{x}, t\right)$ (mol_C/L). The immobile species are the biomass of facultative anaerobic microbes and the natural organic matter (NOM), which serves as sole electron donor.

The reaction rates in Equation 7 to Equation 14 are adapted from Sanz-Prat et al. (2015) and Loschko et al. (2018). The degradation rates r (mol/[ls]) of oxygen and nitrate (Equations 7 and 9) are modeled by standard dual-Monod kinetics (Equations 8 and 10) (Sanz-Prat et al., 2015). Denitrification is inhibited by dissolved oxygen, which is modeled by the noncompetitive inhibition term in Equation 10.

$$r_{O_2} = -\frac{\mu_{O_2}}{Y_{O_2}} \tag{7}$$

$$\mu_{O_2} = \mu_{O_2}^{max} \cdot \frac{c_{O_2}}{c_{O_2} + K_{O_2}} \cdot \frac{c_{DOC}}{c_{DOC} + K_{DOC}} \cdot c_{bac}^* \tag{8}$$

$$r_{NO_3^-} = -\frac{\mu_{NO_3^-}}{Y_{NO_3^-}} \tag{9}$$

$$\mu_{NO_3^-} = \mu_{NO_3^-}^{max} \cdot \frac{c_{NO_3^-}}{c_{NO_3^-} + K_{NO_3^-}} \cdot \frac{c_{DOC}}{c_{DOC} + K_{DOC}} \cdot \frac{K_{O_2}^{inh}}{c_{O_2} + K_{O_2}^{inh}} \cdot c_{bac}^*. \tag{10}$$

Here, μ_i (1/s) is the specific growth rate of component i and Y_i ($\mathrm{mol}_C/\mathrm{mol}_i$) is the yield coefficient. K_i (mol/L) is the Monod constant of the species i and $K_{O_2}^{inh}$ (mol/L) is the inhibition constant of oxygen in denitrification.

The reaction rate of DOC, r_{DOC} (mol_C/[ls]), and the rate of its release from the soil matrix, $r_{DOC}^{rel}(\mathbf{x}, t)$ (mol_C/[ls]), are given in Equations 11 and 12, respectively. To model the release of DOC from natural organic matter (NOM) of the aquifer matrix, we choose a linear driving-force-expression. We assume an infinite supply of NOM; that is, the long-term depletion of NOM is neglected because this process usually takes decades (Loschko et al., 2018; 2019).

$$r_{DOC} = r_{DOC}^{rel} - \left(\frac{\mu_{O_2}}{Y_{O_2}} + \frac{5}{4} \frac{\mu_{NO_3^-}}{Y_{NO_3^-}} \right) \tag{11}$$

$$r_{DOC}^{rel} = k_{DOC}^{rel} \cdot \left(c_{DOC}^{sat} - c_{DOC} \right). \tag{12}$$

The parameter k_{DOC}^{rel} (1/s) is the maximal release rate of DOC from NOM. The reaction rate r_{bac} of the immobile biomass is described in Equation 13 with a decay term r_{dec} given in Equation 14. In Equation 13, c_{bio}^{max} (mol_C/L) is the maximum biomass concentration that accounts for a limited carrying capacity. Biomass decay is modeled using first-order decay with rate coefficient k_{dec} (1/s). It is assumed that the biomass concentration does not fall below a minimum concentration c_{bac}^{min} (mol_C/s).

$$r_{bac} = \left(\mu_{oxy} + \mu_{nit} \right) \cdot \left(1 - \frac{c_{bac}^*}{c_{bac}^{max}} \right) - r_{dec} \tag{13}$$

$$r_{dec} = k_{dec} \cdot \left(c_{bac} - c_{bac}^{min} \right). \tag{14}$$

Solute transport (Equation 5) and reactions (Equations 6–14) are solved together by a fully implicit coupling scheme using Newton-Raphson iteration with adaptive time stepping. This yields the concentrations of all compounds dependent on location **x** and time *t*.

The key assumptions of the reference model are as follows:

1. The NOM concentration is considered constant in time, neglecting the potential decrease of the soil's reaction potential. This assumption seems to be justifiable if the considered processes act on short time scales compared to the depletion of the reaction partners in the soil matrix. According to Loschko et al. (2018), this depletion was observed in aquifers after several years of ongoing denitrification.
2. Biomass is considered immobile. This means that transport of bacteria, including attachment, detachment, straining, and motility, is neglected.
3. The entire biomass participating in the reactions of the dissolved compounds is summarized into a single species.
4. Reaction intermediates are not considered.
5. DOC is treated like a defined species with constant properties.
6. The hydraulic conductivity field is known.

Based on the reference Model M1, we follow two different branches for simplification: neglecting dispersion (M2) and neglecting biomass growth and decay (M3).

3.2. Model M2: Neglecting Dispersion

Model M2 has the same conceptual basis as the reference Model M1. However, in contrast to Model M1, dispersion is neglected. Thus, Equation 5 simplifies to

$$\frac{\partial c_i}{\partial t} + \mathbf{v} \cdot \nabla c_i = r_i \left(\mathbf{c} \left(\mathbf{x}, t \right), \mathbf{x}, t \right). \tag{15}$$

For substances that are introduced diffusively and react with the soil matrix, it is often assumed that dispersive mixing has a minor influence, especially if we are interested in an integral quantity such as the concentration in a groundwater well (e.g., Loschko et al., 2016; 2018). This is typically the case for nitrate of agricultural origin, which is distributed over a relatively large surface area. In contrast, neglected dispersion would be inappropriate for point-like sources such as a contamination plume from a leakage, or when considering the dynamics of an invasion front (e.g., Cirpka et al., 2012).

In our later analyses and discussions, we consider two versions of Model M2 to reflect potential differences in the way that different modelers approach model simplification: In case of Model M2a, we use the same parameter distributions as for Model M1. However, a modeler might decide to modify these parameters to compensate for the effects caused by neglecting dispersion. Therefore, in the second scenario M2b, we shift the prior distribution of the maximum specific growth rates toward higher values and use a log-uniform distribution.

3.3. Model M3: Neglecting Biomass Growth and Decay

Model M3 is based on the same spatially explicit description as the Models M1 and M2, but it neglects the growth and decay of biomass. This means that the biomass concentration remains at its initial value and that Equation 13 simplifies to $r_{bac} = 0$. As a consequence, the solute concentrations and the release of DOC from the soil matrix are the only variables that affect the reaction rates of nitrate, oxygen, and DOC (Equations 7–11). As constant biomass concentration we choose the maximum biomass concentration from Models M1 and M2 as an upper limit and sample the parameter from a uniform distribution between 70% and 100% of the value used in the Models M1 and M2.

The underlying assumption of this simplified model is that typical time scales for the establishment of the microbial community are much smaller than the time scales over which nitrate is introduced into groundwater.

3.4. Model M4: Cumulative Relative Reactivity With Monod Kinetics

Models M4 and M5 follow a completely different approach compared to the aforementioned ones. These models are based on the concept of advective travel times and cumulative relative reactivity. The main idea is to follow the path of a water parcel through the aquifer. Along its trajectory, the water parcel is exposed to geological zones of varying reactivity. The method assumes that the variability in reactivity can be expressed as a fixed ratio of the rate of a chemical reaction for given concentrations to a reference reaction rate at the same concentrations. This relative reactivity is integrated over the travel time of the water parcel, yielding the cumulative relative reactivity. This quantity replaces time in the ordinary differential equations (ODEs) describing the reaction rates of the solutes. In the final step, the concentrations at each location and time can be determined by mapping from cumulative relative reactivity to space. In the following section, the concept is briefly outlined and we refer the reader to Loschko et al. (2016) for a detailed derivation and testing of the concept.

Within the concept of cumulative relative reactivity, the reaction rates in Equation 5 are split up into two parts:

$$\mathbf{r}\left(\mathbf{c}\left(\mathbf{x},t\right),\mathbf{x},t\right)=f\left(\mathbf{x}\right)\mathbf{r}_0\left(\mathbf{c}\left(\mathbf{x},t\right)\right),\tag{16}$$

where $\mathbf{r}_0(\mathbf{c}(\mathbf{x},t))$ (mol/[ls]) is the concentration-dependent reference reaction rate and $f(\mathbf{x})$ (-) is the relative reactivity, which is a concentration-independent, spatially variable scalar multiplier (Loschko et al., 2016). In the given application, the dimensionless relative reactivity $f\left(\mathbf{x}\right)$ accounts for the existence and strength of an electron donor in the soil matrix and specifies the intensity of the considered reaction compared to the reference reaction rate $\mathbf{r}_0\left(\mathbf{c}\left(\mathbf{x},t\right)\right)$. Thus, $f\left(\mathbf{x}\right)$ is directly related to the concentration of NOM, which is assumed to remain at quasi steady state. Loschko et al. (2016) give an example for the interpretation of the relative reactivity and the reference reaction rate: There are three factors influencing the reaction rate of oxygen at a given location in space and time, (1) the oxygen concentration, (2) the nitrate concentration, and (3) the availability of a reaction partner in the matrix. The reference reaction rate $\mathbf{r}_0\left(\mathbf{c}\left(\mathbf{x},t\right)\right)$ includes the first two, whereas the third belongs to the relative reactivity $f\left(\mathbf{x}\right)$.

In the Lagrangian perspective, a water parcel is traced through the domain. Under steady state flow conditions, the position $\mathbf{x}$ of such a parcel depends on its starting location $\mathbf{x}_0$, its velocity $\mathbf{v}(\mathbf{x},t)$ and its travel time τ. Thus,

$$\mathbf{x}\left(\tau|\mathbf{x}_0\right)=\mathbf{x}_0+\int_0^{\tau}\mathbf{v}\left(\mathbf{x}\left(\tau_*|\mathbf{x}_0\right)\right)d\tau_*.\tag{17}$$

Analogously, the cumulative relative reactivity F (T) can be defined as the integral of the relative reactivity $f\left(\mathbf{x}\right)$ along the travel time and therefore as a measure of how long this parcel has been exposed to regions of strong reactivity:

$$F\left(\mathbf{x}\right)=F\left(\tau\left(\mathbf{x}\right)|\mathbf{x}_0\left(\mathbf{x}\right)\right)=\int_0^{\tau}f\left(\mathbf{x}\left(\tau_*|\mathbf{x}_0\right)\right)d\tau_*.\tag{18}$$

Combining Equations 15, 16, and 18 yields

$$\begin{aligned}\frac{d\mathbf{c}}{dF}&=\mathbf{r}_0\left(\mathbf{c}\right)\\ \mathbf{c}\left(t_0\right)&=\mathbf{c}\left(\mathbf{x}_0,t_0\right).\end{aligned}\tag{19}$$

The concentrations of the solutes can be obtained by solving the system of ODEs in Equation 19 and mapping it to the spatial domain. The mapping is defined by the origin, travel time, and cumulative relative reactivity of a water parcel:

$$\mathbf{c}\left(\mathbf{x},t\right)=\mathbf{c}_{ODE}\left(F\left(\mathbf{x},t\right),\mathbf{c}_0\left(\mathbf{x}_0\left(\mathbf{x},t\right),t-\tau\left(\mathbf{x},t\right)\right)\right).\tag{20}$$

This approach reduces computation times tremendously compared to the spatially explicit models (cf. Table 2).

In Model M4, the aerobic respiration and denitrification are described by standard Monod kinetics with noncompetitive inhibition of denitrification by dissolved oxygen:

$$r_{0,O_2}=\frac{dc_{O_2}}{dF}=-\frac{c_{O_2}}{c_{O_2}+K_{O_2}}\cdot r_{max}^{O_2}\tag{21}$$

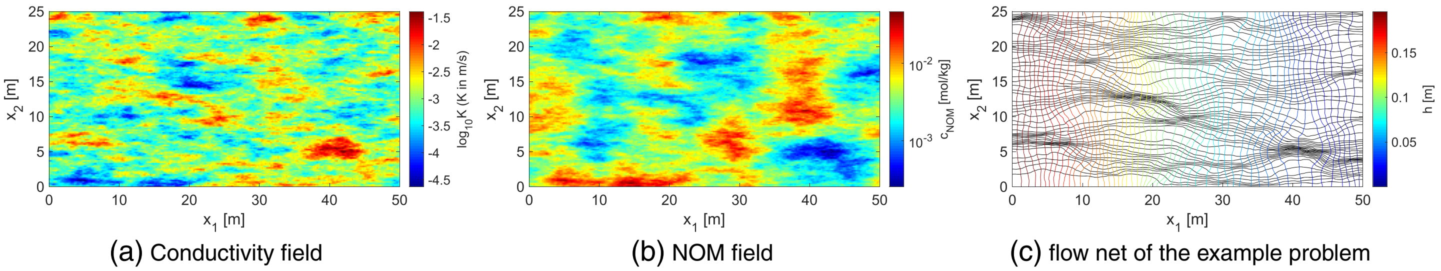


Figure 2. (a) Conductivity field, (b) NOM field, and (c) flow net of the example problem.

$$r_{0,NO_3^-} = \frac{dc_{NO_3^-}}{dF} = -\frac{c_{NO_3^-}}{c_{NO_3^-} + K_{NO_3^-}} \cdot \frac{K_{O_2}^{inh}}{c_{O_2} + K_{O_2}^{inh}} \cdot r_{max}^{NO_3^-}, \tag{22}$$

in which r_{max}^{i} (mol/[ls]) is the maximum reaction rate of the dissolved species i for a relative reactivity of $f(\mathbf{x}) = 1$.

3.5. Model M5: Cumulative Relative Reactivity With Simplified Kinetics

Model M5 has the same conceptual basis as Model M4, but it uses simplified reaction kinetics. Typically, the Monod coefficients K_{O_2} and $K_{O_2}^{inh}$ are relatively small, whereas $K_{NO_3^-}$ is comparatively large (Loschko et al., 2016). These assumptions simplify the Monod terms in Equations 21 and 22 to zeroth-order decay for aerobic respiration and first-order decay for denitrification:

$$r_{0,O_2} = \begin{cases} -r_{max}^{O_2} & \text{if } c_{O_2} > 0 \\ 0 & \text{else} \end{cases} \tag{23}$$

$$r_{0,NO_3^-} = \begin{cases} 0 & \text{if } c_{O_2} > 0 \\ -k_{NO_3^-} \cdot c_{NO_3^-} & \text{else,} \end{cases} \tag{24}$$

in which $k_{NO_3^-}$ (1/s) is the first-order decay coefficient of nitrate. While in Model M4 denitrification is only inhibited when oxygen is available, it is completely prohibited in Model M5. Note that the ODE system of Equation 19 with the rate laws of Equations 23 and 24 has a simple analytical solution.

4. Setup and Implementation

The scenario considered in this study consists of a two-dimensional rectangular domain of size 50 m × 25 m with a numerical grid spacing of 0.2 m in each direction. For the flow field, we generate a multi-Gaussian random field with an exponential covariance function and correlation lengths of 4 m × 1 m using the spectral method of Dietrich and Newsam (1997). The geometric mean of the conductivity is set to $K_g = 10^{-3}$ m/s and the variance of the log-hydraulic conductivity is $\sigma_{lnK}^2 = 1$. The flow field is obtained by solving the groundwater flow equation with fixed-head boundary conditions at the left and right boundaries and no-flow conditions at the top and bottom boundaries on this parameter field. For the relative reactivity field, we assume anticorrelation of NOM content and hydraulic conductivity on a larger scale, because areas with low hydraulic conductivity tend to have a high NOM content (Loschko et al., 2016), while the respective small-scale deviations are uncorrelated. Figure 2 shows (a) the spatial distribution of the log-hydraulic conductivity, (b) the NOM field, and (c) the streamline-oriented grid. All geometrical, geostatistical, hydraulic, and transport parameters are listed in Table A1.

Water with dissolved oxygen ($c_{O_2}^{inf} = 2.5 \cdot 10^{-4}$ mol/L) and nitrate ($c_{NO_3^-}^{inf} = 10^{-4}$ mol/L) infiltrates the system from the left. The inflow concentrations are constant over time. No-flow boundary conditions are assigned to the top and the bottom boundaries, at the left and the right boundary the hydraulic head is fixed. The head difference of 0.2 m leads to a moderate average velocity of 0.4 m/day. Initially, nitrate and oxygen are absent in the domain. For the spatially explicit models (M1, M2, and M3), the initial concentrations of DOC and biomass are set to the saturation concentration of DOC ($c_{DOC}^{ini} = 3 \cdot 10^{-4}$ mol/L) and the maximal biomass concentration ($c_{bac}^{ini} = 83$ μmol/L). The initial and boundary conditions are summarized in Table A2.

For the Bayesian model analysis, we sample the parameters from their prior distributions, which are given in Table A4 for the spatially explicit models and in Table A5for the cumulative relative reactivity models. We define the likelihood function $p\left(\mathbf{y}_0|M_i, \mathbf{u}_{ik}\right)$ as a Gaussian distribution with mean $\mu = 0$. For the measurement error, we assume a standard deviation of $\sigma_{meas} = 10^{-5}$ mol/L for each individual measurement (concentration in one streamtube). As we consider the normalized concentration $c = c/c_{inflow}$, we also normalize the standard deviation $\sigma = \sigma_{meas}/c_{inflow} = 0.1$ mol/L. Our quantity of interest is the normalized concentration averaged over all n_{st} streamtubes at a certain cross section. Because the concentrations in adjacent streamtubes are correlated, we have to account for the correlation of the measurement error variances: We assume equal variance $\sigma^2(c) = 0.01$ mol^2/L^2 for the concentration in each streamtube and calculate the correlation ρ between the concentrations in each streamtube at a certain cross section. The variance of the mean can be calculated as $\sigma^2\left(\bar{c}\right) = \frac{\sigma^2(c)}{n_{st}} + \frac{n_{st}-1}{n_{st}}\rho\sigma^2(c)$. This results in a slightly decreased variance related to the measurement error of the mean concentration $\sigma^2\left(\bar{c}\right) = 0.0092$ mol^2/L^2.

We choose uniform prior model weights $P\left(M_i\right) = 1/N_m$, which means all models are considered as equally likely before seeing any reference data set.

4.1. Numerical Methods

4.1.1. Reactive Transport Models

Heads and stream function are solved by the Finite Element Method with bilinear elements. In the next step, a streamline-oriented grid (Cirpka et al., 1999a) is generated with $n_{st} = 125$ streamtubes and $n_{sec} = 250$ streamtube sections. Figure 2c shows the resulting flow net. We compute the mean groundwater age (Goode, 1996) and cumulative relative reactivity (Equations 17 and 18) along the streamtubes. In Models M1 to M3, advective-dispersive-reactive transport is solved by cell-centered Finite Volumes on the streamline-oriented grid (Cirpka et al., 1999b). Reactions and transport are coupled with a fully implicit scheme using Newton-Raphson iteration with adaptive time stepping, as already done by Loschko et al. (2016).

4.1.2. Bayesian Model Justifiability Analysis

We calculate the BME values by averaging the likelihood values obtained from $N_{MC} = 10^4$ Monte Carlo samples drawn from the parameter priors (Equation 3). The convergence of the BME values is checked by observing that the values reach a steady state over increasing sample size and by determining the effective sample size (ESS) (Liu, 2004). The ESS indicates how many realizations contribute significantly to the BME estimate (Schöniger, Illman, et al., 2015). The ESS values range from 458 for M1 to 2866 for M3 and are hence comfortably high to ensure stable results.

The quantity of interest used for the justifiability analysis is the normalized nitrate concentration, averaged over 125 streamtubes at N_{cs} cross sections. The number of cross sections considered is varied between one and 150 cross sections. Remember that we can afford arbitrarily large data sets, because we work in a synthetic setting and data set size is only limited by grid resolution. The very high resolution data sets do not serve to mimic realistic field conditions but to test and compare the models against each other on a detailed grid.

5. Results and Discussion

The normalized concentrations c/c_0 of nitrate at different cross sections predicted by the six models are shown in Figure 3. The values are averaged over all streamlines of the respective cross section. Based on this figure, we want to analyze (1) how similar the models are independent of the parameter choice, that is, before they are calibrated, and (2) how well the simplified Models M2–M5 can reproduce a specific reference data set that was generated by the reference Model M1. This reference data set is based on a single realization using parameters that are a typical expert choice (cf. Table A3) and is shown as a black line in Figure 3. All shaded areas illustrate the 90% credible intervals of the model predictions in three different states: The light gray intervals show the prior predictions, that is, the models' behavior over the range of parameters that was considered plausible before they are conditioned on data. The prior means are shown as gray lines. The red intervals show the posterior predictions after the models have been calibrated on the flux-weighted nitrate concentration in a single cross section at the outflow boundary. The red lines represent the corresponding posterior means. The dark gray intervals illustrate the posterior predictions after the models have been calibrated on the reference values at 150 equidistant cross sections. The corresponding posterior ensemble means are shown as green lines.

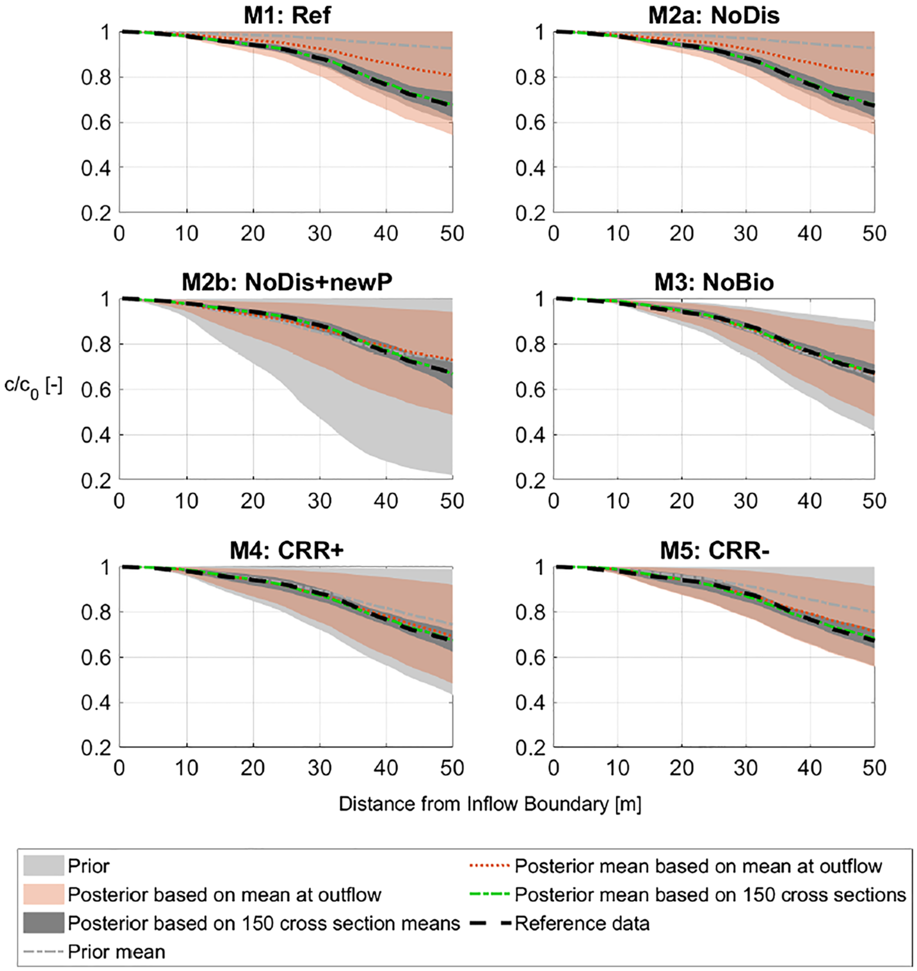


Figure 3. Ninety percent credible intervals for prior state (light gray), posterior after calibrating on one data point at the outflow (red), and posterior after calibrating on 150 equidistant data points (dark gray). Prior means (light gray line), posterior means after calibrating on one data point at the outflow (red line), and posterior means after calibrating on 150 equidistant data points (green line). The reference data (black line) is an arbitrarily chosen realization of M1.

5.1. Prior Predictive Distributions

The analysis of the prior intervals shows that the two scenarios considered for Model M2 differ substantially. While the interval predicted by M2a is remarkably similar to the one predicted by the reference Model M1, M2b has a much higher variance than all other models. This is caused by the attempt to compensate for a neglected process by changing the parameter prior distributions in Scenario 2b. From the predictions made by Model M2a, we can conclude that dispersion is not a dominating process for the considered quantity of interest in the tested setup and therefore it can be neglected without any compensation.

From the prior predictions of M3, it can be seen that the model predicts rather low nitrate concentrations (i.e., high nitrate depletion) and cannot reproduce the cases with very little or even no nitrate depletion in M1. Again, this is an issue related to compensation mechanisms in simplified models: When biomass dynamics are neglected, the biomass concentration is fixed at its initial value. In our setup, this value is sampled from a uniform distribution ranging between 70% and 100% of the maximum biomass concentration in the reference model. However, this choice tends to cause higher nitrate depletion than the reference model. A closer analysis of the reference Model M1 reveals that the realizations with very little nitrate reduction are characterized by high decay coefficients of bacteria (k_{dec}) and low values of the maximum specific growth rates ($\mu_{max}^{O_2}$, $\mu_{max}^{NO_3^-}$). These cases cannot be reproduced by Model M3.

The most simplified Model M5 has a similar prior predictive range as the reference Model M1, while the predictions of M4 show a higher variance than Model M1, though it has only 2 parameters while M1 has 10.

If we think about model complexity only in terms of parametric complexity or "number of parameters included," this result might be surprising. However, the more we simplify a model, the more "effective" (as opposed to mechanistic) its process description becomes. This in turn makes it more difficult to define reasonable prior ranges of the effective parameters, as they lose their physical meaning. This effect becomes clearly visible in the prior predictions of the Models M2b, M3, and M4.

The prior mean of predictions by M1 yields relatively high nitrate concentrations (more than 90% of the initial nitrate concentration reaches the outflow boundary). This is because the predictive distribution of M1 is strongly skewed toward high nitrate concentrations. The same holds for the prior mean of Model M2a. The prior means of M2b to M4 range between 65% and 75% of the inflow concentration at the outflow boundary, while M5 also predicts slightly higher concentrations on average (80% of the initial nitrate concentration reaches the outflow boundary).

5.2. Posterior Predictive Distributions

The posterior credible intervals for the case when only the concentration at the outflow boundary was used for conditioning the models (red intervals) are relatively similar for M2b–M5. The reference Model M1 still covers the range of little nitrate depletion (remember that also M1 was calibrated on its own reference data set), with the highly similar Model M2a showing the same behavior. The posterior means (red line) of M2b–M5 reproduce the reference data quite accurately. Interestingly, the data generating Model M1 performs worse than the simplified Models M2b–M5. The reason for this is the very high exceedance probability of the reference data set (black line) in the predictive distribution of M1. This might be surprising as the parameters that were chosen for the reference data set are mostly located centrally in the prescribed range (cf. Tables A3 and A4). However, the nonlinearity of the simulated processes leads to a highly skewed predictive distribution.

Using 150 cross sections as calibration data leads to a considerable shrinkage of the posterior intervals (dark gray) for all models. In this case, the posterior mean (green line) of all models reproduces the reference data set accurately.

In summary, Figure 3 implies that the Models M2a, M4, and M5 are suitable simplifications of the reference model if quasi steady state concentrations are considered. However, the analysis so far did not take the models' complexity into account. This will be done by the model justifiability analysis, presented in the next section. The prior credible intervals of M2b show that a modeler's uncertainty about compensation mechanisms might lead to overly wide prior choices. For M3, a similar issue became evident: the assumption about the biomass concentration led to higher nitrate depletion than in the reference model. Of course, the cumulative relative reactivity models (M4 and M5) also involve effective parameters. Yet, these models have less parameters than M2 and M3, which makes their predictive distributions less prone to difficulties in the prior formulation.

5.3. Model Justifiability Analysis

All conclusions drawn from the interpretation of the posterior distributions in Figure 3 are conditional on the realization of M1 that was chosen as reference data. To gain a more comprehensive understanding of how suitable Models M2 to M5 are as simplifications of the reference Model M1, we have to consider the overall prediction space of the reference Model M1 instead of picking just a single, somewhat arbitrary realization as reference data set. The model justifiability analysis as described in section 2.2 with M1 as the data generating model fulfills exactly this task. The resulting model weights allow statements about the overall suitability of the simplified models, integrated over the range of data sets that are plausible according to the reference Model M1. Here, suitability means how well the models score in a trade-off between goodness-of-fit to reference data and parsimony.

Figure 4a shows the weights each model receives when the reference Model M1 has generated the data. Starting from equal prior weights of $P\left(M_i\right) = 0.2$, the weights of Models M1 and M2a are very similar and increase monotonically. This confirms the conclusion drawn from Figure 3 that M2a is highly similar to M1. Beyond that, the high weights for Model M2a show that, even for the smallest data set, it scores well independent of the parameter choice and in terms of the tradeoff between model complexity and goodness-of-fit. With increasing data set size, the weights for M3 decrease monotonically. This means that with more data, the dissimilarity between M1 and M3 becomes more evident. The weights for M4 and M5 decrease only at a very slow rate. This shows that, even with the largest data set, there is "confusion" among these models.

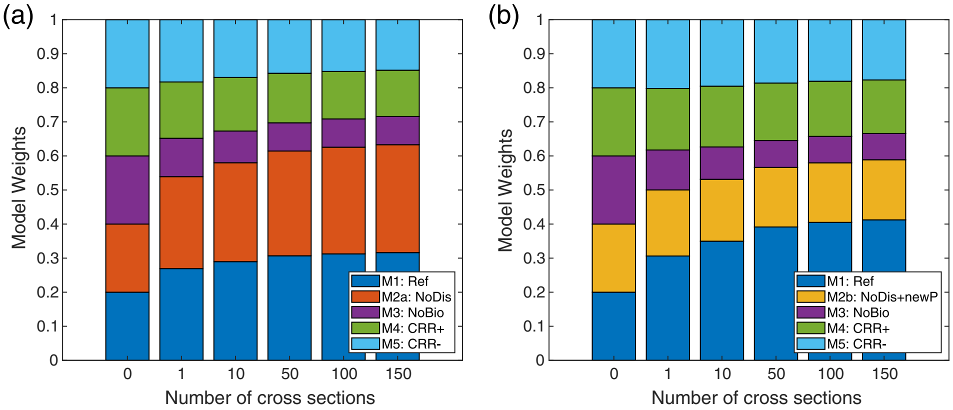
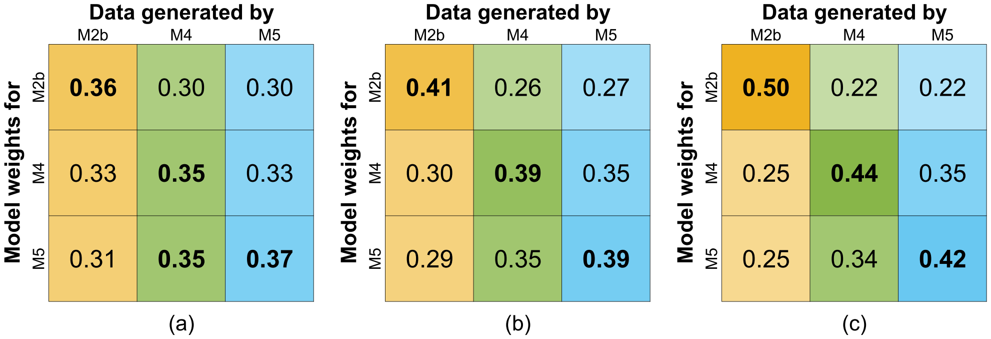


Figure 4. Average model weights over increasing data set size, when M1 generates the data. (a) Using the same parameter distributions for M1 and M2. (b) Testing parameter distributions that compensate for the missing dispersion in M2.

Figure 4b also shows the weights each model receives, when the reference Model M1 has generated the data, but this time replacing Model M2a with its changed-prior variant M2b. As in Scenario A, Model M3 scores worst. Model M2b receives much smaller weights than M2a, because the wide prior is now penalized as overly complex through the eyes of BMS. The weight of M2b is similar to those of M4 and M5. For this modeling scenario, the analyst would now have to decide whether all three models are similarly good candidates to replace M1. To this end, we investigate the similarity between these three candidate models by constructing a 3 × 3 model confusion matrix. We expect that Models M4 and M5 are actually very similar, as they are based on the same modeling concept. From conceptual considerations and the output distributions in Figure 3, we know that Model M2b differs significantly from M4 and M5. The similarity analysis based on the model confusion matrix will now reveal whether the differences are large enough for model discrimination via BMS. Further, we can learn from this analysis whether M2b scores a similar tradeoff between performance and parsimony as M4 and M5. If it did, we would see similar weights for all three models in the off-diagonal entries of the model confusion matrices; here, however, we expect to see a punishment of the complexity (wide prior distribution) of M2b instead, leading to clearly lower weights for M2b if M4 or M5 generate the data.

Figure 5 shows model confusion matrices for increasing data set sizes. The highest weight for each data generating model (column) is printed in bold. For Figure 5a, only the flux-weighted concentration at the outflow boundary was used for the analysis, while Figures 5b and 5c are based on a data set of 25 and 150 cross sections, respectively. The weights on the main-diagonal reflect the ability of each model to identify its own predictions (self-identification weights). The off-diagonal elements are the weights each model receives when the reference data was generated by another model and can be interpreted as a measure of model similarity.


Data generated by
Model weights for

	M2b	M4	M5
M2b	**0.36**	0.30	0.30
M4	0.33	**0.35**	0.33
M5	0.31	**0.35**	**0.37**

(a)

	M2b	M4	M5
M2b	**0.41**	0.26	0.27
M4	0.30	**0.39**	0.35
M5	0.29	0.35	**0.39**

(b)

	M2b	M4	M5
M2b	**0.50**	0.22	0.22
M4	0.25	**0.44**	0.35
M5	0.25	0.34	**0.42**

(c)

Figure 5. Model confusion matrices for M2b, M4, M5 based on concentrations in (a) a single cross section and (b) 25 and (c) 150 cross sections. Columns refer to the models that generate the data, rows to the models that we calculate the weights for. The highest weight for each data generating model (column) is printed in bold.

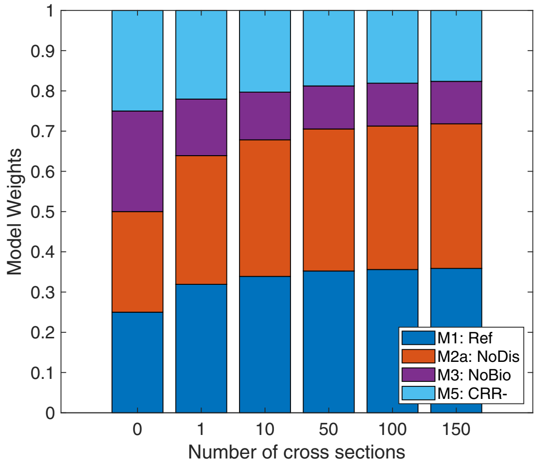


Figure 6. Average model weights when M1 generates the data over increasing data set size for the reduced model set.

When the size of the data set is increased, the self-identification weights of all models increase (see Figures 5a–5c). This agrees with the theoretical expectation that, in the BMS framework, the true model can be identified in the limit of infinite data set size (Schöniger, Illman, et al., 2015). However, in the setting we analyze here, the self-identification weights increase only slowly. This is probably due to correlations between the concentrations in the individual cross sections, so that more data only add limited information with respect to model choice.

In the last two columns (when M4 and M5 generated the data), the "confusion" among the models remains strong: Even with the largest data set size, there is not yet a clear picture of model identification. The off-diagonal elements reveal that M4 and M5 are actually similar. M2, however, does not show such a strong confusion with other models. This confirms our attempt to explain the similar weights of M2b, M4, and M5 when the reference Model M1 generated the data: The confusion matrices imply that M4 and M5 are almost redundant for the purpose of predicting nitrate concentration. M2b instead is different in its predictions but scores a similar tradeoff between goodness-of-fit and parsimony.

5.4. Justifiability Analysis for a Reduced Model Set

From the analysis so far, we can conclude that the predictions of the two cumulative relative reactivity models (M4, M5) are very similar and, consequently, that the models are quasi-redundant for the purpose of predicting nitrate concentration. Therefore, we decide to exclude one of these two models to avoid misinterpretations due to redundancy in the model set. Considering that M5 is slightly preferred over M4 for data sets generated by the reference Model M1 (cf. Figure 4) and that M5 is the most parsimonious model in the set, we decide to keep it and discard M4. Also, the analysis revealed clearly that M2a is a better choice than M2b. Therefore, we omit M2b from the following analysis. With this reduced model set, we want to test how M5 scores compared to the other simplified models, now that it does not have to compete with a very similar model.

Figure 6 shows the weights the models receive when the reference Model M1 has generated the data. Comparison with Figure 4 shows that excluding M4 leads to a redistribution of model weights due to the constraint that they sum up to 1. The strongest relative increase is in fact in M1 and not in M5. Note that, for individual data sets, the relative increase in model weights is the same for all four models, and it can be calculated as $\sum_{i=1}^{5} BME_i / \sum_{i=1}^{4} BME_i$. However, due to averaging over many data sets representing the predictive distribution of M1 and the large variations in BME values per data set, this constant factor translates into individual reweighting factors per model. The more decisive the model weighting, the more nonlinear the behavior.

5.5. Decisiveness of Model Choice Measured by Bayes Factors

We want to further investigate the decisiveness of model choice by using Bayes Factors. To this end, we use the BME values to determine pairwise Bayes factors between the reference model and the simplified models for a data set of 150 cross sections. We ask "how much stronger is the evidence in favor of the reference model M1" and therefore calculate

$$BF = \frac{BME_1}{BME_i}, \text{with } i = 2 - 5. \tag{25}$$

Figure 7 shows the cumulative distributions functions (CDFs) of $log_{10}(BF)$ for each of the simplified models. The dashed lines mark the thresholds according to Jeffreys (1961) (cf. section 2.3). The light gray line ($log_{10}(BF) = 0$) indicates equal support for both models. The black line at $log_{10}(BF) = 2$ indicates "decisive" evidence against the simplified model as an alternative to M1. Vice versa, $log_{10}(BF) = -2$ indicates "decisive" evidence *in favor* of the simplified model, so in these cases the simplified model should be preferred through the eyes of BMS, even though the underlying data set has in fact been generated by the reference M1. Such cases occur because of the complexity of M1: if the simplified model is able to fit the data well and shows less variability than M1, it will be preferred by BMS (see section 1).

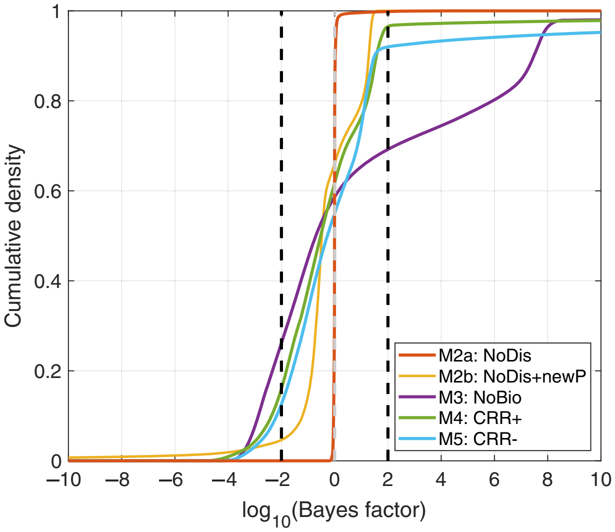


Figure 7. CDFs of the logarithmic Bayes Factor for M2–M5 tested against the reference Model M1.

The CDF of Model M2a exemplifies the distribution for a model that is almost identical to the reference model, as for nearly all realizations $log_{10}(BF) = 0$. The median log_{10} Bayes factors of the other simplified models are all slightly negative (−0.49 for M2b, −0.70 for M3, −0.54 for M4, and −0.28 M5) and thus show a small preference of the simplified models over the reference model. The Bayes Factor CDFs of M2b, M4, and M5 are relatively similar and show that less than about 15% of their realizations lead to a rejection with decisive evidence.

In contrast, M3 shows a high variability in its performance: 30% of its realizations are rejected with "decisive" evidence. At the same time, 26% of its realizations are preferred over the reference model with "decisive" evidence. The reason why the CDF of M3 considerably deviates from the other curves is that, in contrast to all other models, M3 cannot reproduce very high nitrate concentrations. Thus, for realizations of M1 with very little nitrate depletion, M3 has extremely small BME values and consequently, is rejected clearly against the reference Model M1. However, if M3 is able to reproduce the realizations of M1 (i.e., for normalized concentrations less than approximately 0.9), BMS rewards that the predictive distribution of M3 has a higher probability mass concentrated at these values, while the distribution of M1 is strongly skewed toward high nitrate concentrations.

Overall, we find that the CDFs for the Bayes factors against M2b, M4, and M5 support our general conclusion from the average model weights (Figure 4) that all three models are suitable but not perfect candidates to replace the reference Model M1, while M3 is not a robust choice. Model M2a is able to perfectly mimic the reference data. However, through the eyes of BMS, it does not improve in terms of model parsimony; that is, it is still rather complex.

6. Conclusions

In this study, we have a applied the Bayesian model justifiability analysis (Schöniger, Illman, et al., 2015) to compare five models that simulate aerobic respiration and denitrification in a heterogeneous aquifer coupled to solute transport. The model that includes the most detailed description of the underlying processes has served as a reference, whereas the other models were either direct simplifications of the reference model by dropping specific processes, or replaced the advection-(dispersion)-reaction equation in Cartesian coordinates by the concept of cumulative relative reactivity solved along trajectories (Loschko et al., 2016).

The results of the model justifiability analysis show that all simplified models are suitable replacements for the computationally expensive reference model, but the models differ significantly in the number of processes/parameters involved, and in the ease of constructing meaningful prior distributions. In the model justifiability analysis, models are tested against each other based on their prior predictive distributions.

These distributions are generated by sampling from the models' parameter spaces and running the models with the respective parameter realizations in a Monte Carlo framework. By taking the entire predictive distributions into account, the results of the justifiability analysis allow statements about the overall suitability of the simplified models, independent of a specific parameter choice. The analysis is based on the principles of BMS and thus implicitly performs a tradeoff between goodness-of-fit to reference data and model complexity. We highly recommend applying this framework when judging the exchangeability of competing models. However, the process-based reasoning leading to the competing models should never be discarded.

As criterion for model similarity we considered flux-weighted nitrate concentrations in quasi steady state at different cross sections. This choice has direct consequences for the suitability of the model simplifications. Model M2a, which neglects local dispersion and keeps the priors of all other parameters, was practically indistinguishable from the full model. Similar observations have been made by Sanz-Prat et al. (2015), yet without a full stochastic analysis. Local dispersion would have been much more important if we had considered an invading front and a reaction between purely dissolved compounds, or dynamic electron-acceptor loads. Also, biomass dynamics are important predominantly under conditions when the biomass still has to grow, for example, when an aquifer is loaded with a reactant for the very first time. Such conditions have not been considered in the current analysis, as they are rather unlikely for nitrate contamination in aquifers. Because we considered the nitrate concentration after establishment of a stable microbial community, the models without dynamic biomass (i.e., Models M3–M5) had a chance of meeting the reference model. That the spatially explicit model without dynamic biomass scored poorly is mostly due to the difficulty of defining a reasonable prior distribution of the constant biomass concentration and might have been avoided by choosing a broader prior. For the given type of data, the simplest Model M5 using the cumulative-relative-reactivity concept with simplified kinetics turned out to be the best simplification of the computationally expensive reference model. It scores well in the justifiability analysis and reduces run times tremendously compared to the spatially explicit models. This computational efficiency enables a high number of models runs and thus quantification of parametric uncertainty was feasible, which can be impractical with spatially explicit models. As M5 has only two parameters, it is less prone to the problem of overly wide prior ranges. Please note that this recommendation is conditional on the purpose of the model (prediction), the considered scenario (diffusively introduced nitrate reacting with the soil matrix) and the quantity of interest (quasi steady state nitrate concentration as an integral quantity flux-averaged over a cross section).

The present study underpins the currently evolving perception and acknowledgment of complexity in modeling: When we discuss complexity of numerical models, we have to take more into account than the plain number of incorporated processes, interactions and feedbacks. The simplification of physical descriptions often comes at the cost of a more complicated definition of the parameter priors. When we neglect a certain process, the parameters may not represent a physically meaningful value anymore but rather be an effective parameter that compensates for the missing process. Consequently, modelers might encounter difficulties when trying to define realistic prior distributions for effective parameters. The more effective parameters a model has, the stronger this effect can be. Therefore, we emphasize to also consider the constrainability of the parameters as an aspect of model complexity. This means, to take into account how easy or difficult it is to a priori constrain the parameters based on expert knowledge.

We found that performing the justifiability analysis on the case of model simplification is an objective and comprehensive approach to assess the suitability of candidate models with different levels of detail. The method has three major advantages:

- Models are compared independent of calibration data, which might not be available or, as pointed out by Vogel and Sankarasubramanian (2003), even “cloud our ability” to accept or reject a model concept.
- Considering the models' entire parameter and predictive distributions provides a comprehensive model evaluation rather than a comparison based on specific parameter sets.

- Working on the intermodel level, the method allows to filter a set of models with respect to their (prior) predictive power such that, on a second level, a subset of similarly capable models can be rated on additional performance criteria like run time or goodness-of-fit with actual measurement data.

Future research should target model comparison also on the level of structural similarity (Bennett et al., 2019) to complement the analysis of the model predictions. This might help detect structural redundancy in the model set and can further advance directed model (set) development.

Appendix A: Parameters and Initial Conditions

The following tables provide implementation details such as the geometrical, geostatistical, hydraulic and transport parameters (Table A1), initial and boundary conditions (Table A2), parameters of the reference solution (Table A3) and prior distributions chosen for the uncertain parameters (Table A4 and Table A5).

Table A1
Geometrical, Geostatistical, Hydraulic, and Transport Parameters

Symbol	Meaning	Value	Units
L	Length of the 2-D domain	50	(m)
W	Width of the 2-D domain	25	(m)
n_x	Number of cells in x direction	250	(-)
n_y	Number of cells in y direction	125	(-)
Δx	Cell size in x direction	0.2	(m)
Δy	Cell size in y direction	0.2	(m)
n_{st}	Number of streamtubes	125	(-)
n_{sec}	Number of streamtube sections	250	(-)
Geostatistical parameters of the K-field			
l_x	Correlation length in x direction	4	(m)
l_y	Correlation length in y direction	1	(m)
σ^2_{lnK}	Variance of log-hydraulic conductivity	1	(-)
K_g	Geometric mean of hydraulic conductivity	$1 \cdot 10^{-3}$	(m/s)
Parameters of the flow field			
K_{eff}	Effective hydraulic conductivity	$1.2 \cdot 10^{-3}$	(m/s)
$\bar{q}_x$	Mean specific discharge	0.4	(m/day)
J	Mean hydraulic gradient	$4 \cdot 10^{-3}$	(-)
Transport parameters			
θ	Porosity	0.3	(-)
α_l	Longitudinal dispersivity	$1 \cdot 10^{-2}$	(m)
α_t	Transverse dispersivity	$1 \cdot 10^{-3}$	(m)
D_p	Molecular diffusion coefficient	$1 \cdot 10^{-9}$	(m^2/s)

Table A2
Initial and Boundary Conditions

Symbol	Meaning	Initial conc.	Inflow conc.
$c_{mob}^{O_2}$	Dissolved oxygen (mobile phase)	0 mol/L	$2.5 \cdot 10^{-4}$ mol/L
$c_{mob}^{NO_3^-}$	Nitrate (mobile phase)	0 mol/L	$1 \cdot 10^{-4}$ mol/L
$c_{mob}^{CH_2O}$	Dissolved organic carbon (mobile phase)	$3 \cdot 10^{-4}$ mol/L	0 mol/L
c_{immob}^{bac}	Bacteria (immobile phase)	80 µmol/L	n.a.

Table A3
Parameters of the Reference Solution (M1)

Symbol	Meaning	Value
$\mu_{max}^{O_2}$	Maximum specific growth rate based on oxygen	0.1 (1/day)
$\mu_{max}^{NO_3^-}$	Maximum specific growth rate based on nitrate	0.1 (1/day)
K_{O_2}	Monod coeff. of oxygen	11.4 (µmol/L)
$K_{NO_3^-}$	Monod coeff. of nitrate	70 (µmmol/L)
K_{DOC}	Monod coeff. of DOC	20 (µmol/L)
$K_{inh}^{O_2}$	Inhibition coeff. of oxygen in denitrification	10 (µmol/L)
Y_{O_2}	Yield coeff. of oxygen	0.25 ($\mathrm{mol}_{O_2}^{bac}/\mathrm{mol}_C$)
$Y_{NO_3^-}$	Yield coeff. of nitrate	0.25 ($\mathrm{mol}_{NO_3^-}^{bac}/\mathrm{mol}_C$)
k_{dec}	Decay coeff. of bacteria	0.05 (1/day)
$k_{DOC}^{rel,max}$	Maximum rate constant of DOC release	0.2 (1/day)
c_{max}^{bac}	Maximum biomass concentration	83.3 ($\mu\mathrm{mol}_C$/L)

Table A4
The datasets generated and analyzed during the current study are available in the FDAT repository of the University of Tübingen, https://fdat.escience.uni-tuebingen.de/portal/

Symbol	Meaning	Distribution		Units
Parameters of Models M1 and M2a				
$\mu_{max}^{O_2}$	Maximum specific growth rate based on oxygen	$unif.(a,b)$	$a = 1.5 \cdot 10^{-3}$, $b = 0.12$	(1/day)
$\mu_{max}^{NO_3^-}$	Maximum specific growth rate based on nitrate	$unif.(a,b)$	$a = 1.5 \cdot 10^{-3}$, $b = 0.12$	(1/day)
K_{O_2}	Monod coefficient of oxygen	$unif.(a,b)$	$a = 5, b = 15$	(µmol/L)
$k_{NO_3^-}$	Monod coefficient of nitrate	$unif.(a,b)$	$a = 60, b = 80$	(µmol/L)
K_{DOC}	Monod coefficient of DOC	$unif.(a,b)$	$a = 10, b = 30$	(µmol/L)
$K_{inh}^{O_2}$	Inhibition coefficient of oxygen in denitrification	$unif.(a,b)$	$a = 5, b = 15$	(µmol/L)
Y_{O_2}	Yield coefficient of oxygen	$unif.(a,b)$	$a = 0.2, b = 0.3$	($\mathrm{mol}_{O_2}^{bac}/\mathrm{mol}_C$)
$Y_{NO_3^-}$	Yield coefficient of nitrate	$unif.(a,b)$	$a = 0.2, b = 0.3$	($\mathrm{mol}_{NO_3^-}^{bac}/\mathrm{mol}_C$)
k_{dec}	Decay coefficient of bacteria	$unif.(a,b)$	$a = 0.025, b = 0.075$	(1/day)
$k_{DOC}^{rel,max}$	Maximum rate constant of DOC release	$unif.(a,b)$	$a = 0.1, b = 0.5$	(1/day)
Parameters of Model M2b that differ from Model M1 and M2a				
$\mu_{max}^{O_2}$	Maximum specific growth rate based on oxygen	$log - unif.(a,b)$	$a = 0.05, b = 0.2$	(1/day)
$\mu_{max}^{NO_3^-}$	Maximum specific growth rate based on nitrate	$log - unif.(a,b)$	$a = 0.05, b = 0.2$	(1/day)
Parameters of Model M3 that differ from Models M1 and M2				
c_{bio}^{max}	Maximum biomass concentration	$unif.(a,b)$	$a = 58.3, b = 83.3$	(mol_C/L)

Note. The parameters specified for Model M1 are also applied for Models M2a, M2b, and M3. Exceptions are mentioned separately.

Table A5
Prior Distributions Chosen for the Uncertain Parameters of the Models M4 and M5

Parameters of Model M4				
$r_{max}^{O_2}$	Maximum reaction rate of oxygen under reference conditions	$unif\,.(a,b)$	$a = 2 \cdot 10^{-3}$, $b = 100$	(μmol/[L day])
$r_{max}^{NO_3^-}$	Maximum reaction rate of nitrate under reference conditions	$unif\,.(a,b)$	$a = 2 \cdot 10^{-3}$, $b = 5$	(μmol/[L day])
K_{O_2}	Monod constant for oxygen	$unif\,.(a,b)$	$a = 5$, $b = 15$	(μmol/L)
$k_{NO_3^-}$	Monod constant for nitrate	$unif\,.(a,b)$	$a = 60$, $b = 80$	(μmol/L)
$K_{O_2}^{inh}$	Inhibition coefficient of oxygen in denitrification	$unif\,.(a,b)$	$a = 5$, $b = 15$	(μmol/L)
Parameters of Model M5				
$r_{max}^{O_2}$	Maximum reaction rate of oxygen under reference conditions	$unif\,.(a,b)$	$a = 2 \cdot 10^{-3}$, $b = 100$	(μmol/[L day])
$r_{max}^{NO_3^-}$	Maximum reaction rate of nitrate under reference conditions	$unif\,.(a,b)$	$a = 50$, $b = 25 \cdot 10^3$	(μmol/[L day])

Data Availability Statement

Data are currently being archived in the repository of the University of Tübingen (https://fdat.escience.uni-tuebingen.de/) and will be made available upon acceptance. For review purposes, data can be downloaded using the following link (https://bwsyncandshare.kit.edu/s/iRZzriE9kS6B8bX).

Acknowledgments
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) 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.

References

Almasri, M. N., & Kaluarachchi, J. J. (2007). Modeling nitrate contamination of groundwater in agricultural watersheds. *Journal of Hydrology*, *343*(3–4), 211–229. https://doi.org/10.1016/j.jhydrol.2007.06.016

Alpaydin, E. (2004). *Introduction to machine learning*. Adaptive computation and machine learning. MIT Press.

Atchley, A. L., Maxwell, R. M., & Navarre-Sitchler, A. K. (2013). Using streamlines to simulate stochastic reactive transport in heterogeneous aquifers: Kinetic metal release and transport in CO^2 impacted drinking water aquifers. *Advances in Water Resources*, *52*, 93–106. https://doi.org/10.1016/j.advwatres.2012.09.005

Baartman, J. E., Melsen, L. A., Moore, D., & van der Ploeg, M. J. (2020). On the complexity of model complexity: Viewpoints across the geosciences. *Catena*, *186*(104), 261. https://doi.org/10.1016/j.catena.2019.104261

Babu, G. J. (2011). Resampling methods for model fitting and model selection. *Journal of Biopharmaceutical Statistics*, *21*(6), 1177–1186. https://doi.org/10.1080/10543406.2011.607749

Bennett, A., Nijssen, B., Ou, G., Clark, M., & Nearing, G. (2019). Quantifying process connectivity with transfer entropy in hydrologic models. *Water Resources Research*, *55*, 4613–4629. https://doi.org/10.1029/2018WR024555

Bernardo, J. M., Berger, J. O., Dawid, A., & Clyde, M. (1999). Bayesian model averaging and model search strategies.

Brunetti, G., Šimůnek, J., Glöckler, D., & Stumpp, C. (2020). Handling model complexity with parsimony: Numerical analysis of the nitrogen turnover in a controlled aquifer model setup. *Journal of Hydrology*, *584*(124), 681. https://doi.org/10.1016/j.jhydrol.2020.124681

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd). New York: Springer. oCLC: ocm48557578.

Cirpka, O. A., Frind, E. O., & Helmig, R. (1999a). Streamline-oriented grid generation for transport modelling in two-dimensional domains including wells. *Advances in Water Resources*, *22*(7), 697–710.

Cirpka, O. A., Frind, E. O., & Helmig, R. (1999b). Numerical methods for reactive transport on rectangular and streamline-oriented grids. *Advances in Water Resources*, *22*(7), 711–728. https://doi.org/10.1016/S0309-1708(98)00051-7

Cirpka, O. A., Rolle, M., Chiogna, G., de Barros, F. P., & Nowak, W. (2012). Stochastic evaluation of mixing-controlled steady-state plume lengths in two-dimensional heterogeneous domains. *Journal of Contaminant Hydrology*, *138–139*, 22–39. https://doi.org/10.1016/j.jconhyd.2012.05.007

Cremers, K. J. M. (2002). Stock return predictability: A Bayesian model selection perspective. *The Review of Financial Studies*, *15*(4), 27.

Dagan, G., & Nguyen, V. (1989). A comparison of travel time and concentration approaches to modeling transport by groundwater. *Journal of Contaminant Hydrology*, *4*(1), 79–91. https://doi.org/10.1016/0169-7722(89)90027-2

Dietrich, C. R., & Newsam, G. N. (1997). Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, *18*(4), 1088–1107. https://doi.org/10.1137/S1064827592240555

Enemark, T., Peeters, L. J., Mallants, D., Batelaan, O., Valentine, A. P., & Sambridge, M. (2019). Hydrogeological Bayesian hypothesis testing through trans-dimensional sampling of a stochastic water balance model. *Water*, *11*(7), 1463. https://doi.org/10.3390/w11071463

Ferré, T. P. (2017). Revisiting the relationship between data, models, and decision-making. *Groundwater*, *55*(5), 604–614. https://doi.org/10.1111/gwat.12574

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Goode, D. J. (1996). Direct simulation of groundwater age. *Water Resources Research*, *32*(2), 289–296. https://doi.org/10.1029/95WR03401

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

Guthke, A., Höge, M., & Nowak, W. (2017). Bayesian model evidence as a model evaluation metric, *EGU general assembly conference abstracts* (Vol. 19, pp. 13,390). Vienna:). European Geosciences Union.

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

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

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, *85*(1), 3–28. https://doi.org/10.1890/14-0661.1

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64–72.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Clarendon.

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

Lever, J., Krzywinski, M., & Altman, N. (2016). Model selection and overfitting. *Nature Methods*, *13*(9), 703–704. https://doi.org/10.1038/nmeth.3968

Liu, J. S. (2004). *Monte carlo strategies in scientific computing*, Springer series in statistics. New York, NY: Springer. https://doi.org/10.1007/978-0-387-76371-2

Liu, K., Zhu, Y., Ye, M., Yang, J., Cheng, X., & Shi, L. (2018). Numerical simulation and sensitivity analysis for nitrogen dynamics under sewage water irrigation with organic carbon, water. *Air, & Soil Pollution*, *229*(6), 173. https://doi.org/10.1007/s11270-018-3832-z

Loschko, M., Wöhling, T., Rudolph, D. L., & Cirpka, O. A. (2016). Cumulative relative reactivity: A, concept for modeling aquifer-scale reactive transport. *Water Resources Research*, *52*, 8117–8137. https://doi.org/10.1002/2016WR019080

Loschko, M., Wöhling, T., Rudolph, D. L., & Cirpka, O. A. (2018). Accounting for the decreasing, reaction potential of heterogeneous aquifers in a stochastic framework of aquifer-scale reactive transport. *Water Resources Research*, *54*, 442–463. https://doi.org/10.1002/2017WR021645

Loschko, M., Wöhling, T., Rudolph, D. L., & Cirpka, O. A. (2019). An electron-balance based approach to predict the decreasing denitrification potential of an aquifer. *Groundwater*, *57*(6), 925–939. https://doi.org/10.1111/gwat.12876

Nearing, G. S., & Gupta, H. V. (2018). Ensembles vs. information theory: Supporting science under uncertainty. *Frontiers of Earth Science*, *12*(4), 653–660. https://doi.org/10.1007/s11707-018-0709-9

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A. L., & Troldborg, L. (2012). Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Advances in Water Resources*, *36*, 36–50. https://doi.org/10.1016/j.advwatres.2011.04.006

Rojas, R., Feyen, L., & Dassargues, A. (2008). Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research*, *44*, W12418. https://doi.org/10.1029/2008WR006908

Rojas, R., Kahunde, S., Peeters, L., Batelaan, O., Feyen, L., & Dassargues, A. (2010). Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling. *Journal of Hydrology*, *394*(3–4), 416–435. https://doi.org/10.1016/j.jhydrol.2010.09.016

Sanz-Prat, A., Lu, C., Amos, R. T., Finkel, M., Blowes, D. W., & Cirpka, O. A. (2016). Exposure-time based modeling of nonlinear reactive transport in porous media subject to physical and geochemical heterogeneity. *Journal of Contaminant Hydrology*, *192*, 35–49. https://doi.org/10.1016/j.jconhyd.2016.06.002

Sanz-Prat, A., Lu, C., Finkel, M., & Cirpka, O. A. (2015). On the validity of travel-time based nonlinear bioreactive transport models in steady-state flow. *Journal of Contaminant Hydrology*, *175–176*, 26–43. https://doi.org/10.1016/j.jconhyd.2015.02.003

Sanz-Prat, A., Lu, C., Finkel, M., & Cirpka, O. A. (2016). Using travel times to simulate multi-dimensional bioreactive transport in time-periodic flows. *Journal of Contaminant Hydrology*, *187*, 1–17. https://doi.org/10.1016/j.jconhyd.2016.01.005

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

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

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

Steefel, C. I., & Lichtner, P. C. (1998). Multicomponent reactive transport in discrete fractures: I. Controls on reaction front geometry. *Journal of Hydrology*, *209*(1), 186–199.

Troldborg, L., Refsgaard, J. C., Jensen, K. H., & Engesgaard, P. (2007). The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. *Hydrogeology Journal*, *15*(5), 843–860. https://doi.org/10.1007/s10040-007-0192-y

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*(0), 142–228. https://doi.org/10.1214/12-SS102

Vogel, R. M., & Sankarasubramanian, A. (2003). Validation of a watershed model without calibration. *Water Resources Research*, *39*(10), 1292. https://doi.org/10.1029/2002WR001940

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107.

Zhang, H., Yang, R., Guo, S., & Li, Q. (2020). Modeling fertilization impacts on nitrate leaching and groundwater contamination with HYDRUS-1D and MT3DMS. *Paddy and Water Environment*, *18*, 481–498. https://doi.org/10.1007/s10333-020-00796-6