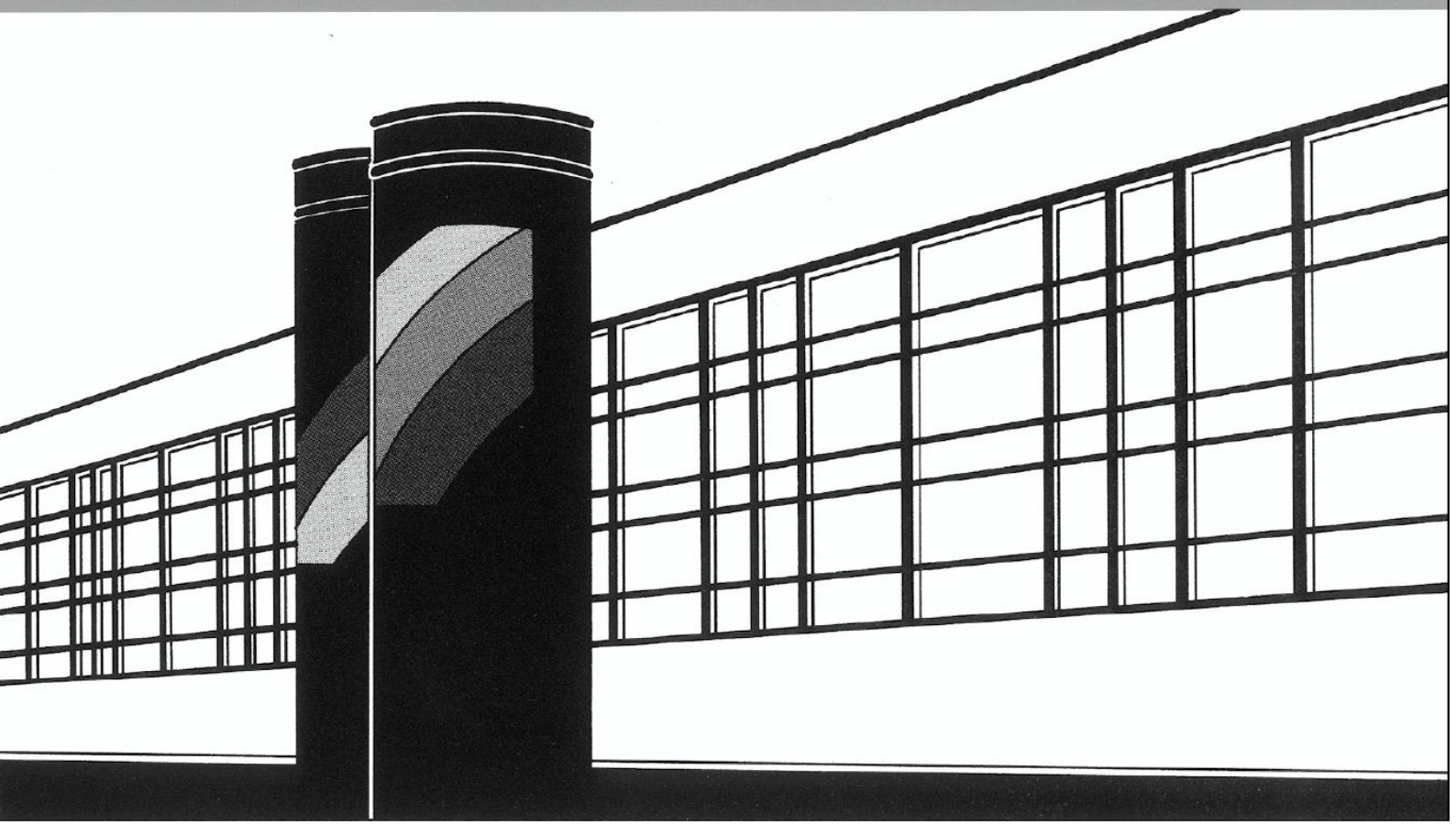
Universität Stuttgart

Institut für Wasser- und Umweltsystemmodellierung

Mitteilungen



Heft 290 Aline Schäfer Rodrigues Silva

Quantifying and Visualizing Model Similarities for Multi-Model Methods

Quantifying and Visualizing Model Similarities for Multi-Model Methods

von der Fakultät Bau- und Umweltingenieurwissenschaften der Universität Stuttgart und dem Stuttgarter Zentrum für Simulationswissenschaften zur Erlangung der Würde eines Doktor-Ingenieurs (Dr.-Ing.) genehmigte Abhandlung

vorgelegt von

Aline Schäfer Rodrigues Silva

aus Ludwigsburg

Hauptberichter:	Prof. Dr.-Ing. Wolfgang Nowak
Mitberichter:	Prof. Ty P.A. Ferré, Ph.D.
Mitberichterin:	Dr. Dipl.-Ing. Anneli Guthke

Tag der mündlichen Prüfung: 26. April 2022

Institut für Wasser- und Umweltsystemmodellierung
der Universität Stuttgart
2022

Heft 290

Quantifying and Visualizing Model Similarities for Multi-Model Methods

von
Dr.-Ing.
Aline Schäfer Rodrigues Silva

Eigenverlag des Instituts für Wasser- und Umweltsystemmodellierung
der Universität Stuttgart

D93 Quantifying and Visualizing Model Similarities for Multi-Model Methods

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://www.d-nb.de abrufbar

Schäfer Rodrigues Silva, Aline:
Quantifying and Visualizing Model Similarities for Multi-Model Methods, Universität Stuttgart. - Stuttgart: Institut für Wasser- und Umweltsystemmodellierung, 2022

(Mitteilungen Institut für Wasser- und Umweltsystemmodellierung, Universität Stuttgart: H. 290)
Zugl.: Stuttgart, Univ., Diss., 2022
ISBN 978-3-942036-94-8

NE: Institut für Wasser- und Umweltsystemmodellierung <Stuttgart>: Mitteilungen

Herausgegeben 2022 vom Eigenverlag des Instituts für Wasser- und Umweltsystem-modellierung

Druck: DCC Kästl e.K., Ostfildern

Für Oma.

Danksagung / Acknowledgements

Ich möchte folgenden Menschen ganz herzlich dafür danken, dass sie mich während dieser Promotion begleitet und unterstützt haben:

Wolfgang: Danke, dass du mir diese große Chance und viel Freiheit gegeben hast. Aus unseren Diskussionen habe ich sehr viel mitgenommen - manchmal auch die Erfahrung, was es bedeutet „gebrainstormt“ zu werden. Gleichermaßen wird mir unser Fachsimpeln über Musik, deine lockere Art und dein Humor in bester Erinnerung bleiben.

Anneli: Du hast oft die Struktur zurückgebracht, wenn bei mir Verwirrung herrschte, und warst eine großartige Lehrmeisterin. Mit deinem Enthusiasmus für die Forschung konntest du mich immer mitreißen und bei all der Arbeit gab es mit dir immer was zu Lachen. Vielen herzlichen Dank für deine Unterstützung!

Ty: Thank you so much for your enthusiasm and support of this project – and for your good humor, even when you got dragged into an exam early in the morning.

Ute: Ein herzliches Dankeschön dafür, dass du immer ein offenes Ohr und einen guten Rat für alle Lebenslagen hattest. Gleiches gilt für deine Geduld und Hilfe, wenn ich mal wieder kein glückliches Händchen für Formulare hatte.

Marvin: Du warst ein großartiger Diskussionspartner für fachliche, politische und viele weitere spannende Themen. Dein Rat war immer Gold wert – herzlichen Dank dafür!

Sebastian: Vielen Dank für den gemeinsamen Denksport, dein immer ehrliches, wertvolles Feedback und jede Menge Spaß auf unseren gemeinsamen Dienstreisen.

Farid: Du warst eine großartige Reisebegleitung durch die Welt des surrogate modelings und auf Korsika! Vielen Dank für deine große Hilfsbereitschaft und das Korrekturlesen dieser Arbeit.

Sergey: Du hast jeden Tag gute Laune ins Büro gebracht und mit unendlicher Geduld bei sämtlichen Fragen weitergeholfen. Herzlichen Dank, es war eine Freude mit dir zu Arbeiten!

Julia, Simón und Micha: Ihr wart wunderbare ZimmergenossInnen und hattet immer ein offenes Ohr. Vielen Dank für die Gespräche über den Tellerrand, fürs Süßigkeiten teilen und für den vielen Spaß.

Jannik: Danke, dass du mir beigebracht hast, auch mal pragmatisch zu sein und immer für gute Laune gesorgt hast.

Timothy: Ein großes Dankeschön für die technische Unterstützung bei der Verteidigung und die gute Laune, die du immer mit zur Arbeit gebracht hast.

Meinen StudentInnen Stefania, Charlotte, Anas, Hanna und Timo: Ihr habt einen großen Beitrag zum Gelingen dieser Arbeit geleistet und wart eine wertvolle Unterstützung in der Lehre. Ein großes Dankeschön dafür!

Meinen Co-Autoren Toby, Sebastian, Thilo, Johannes, Bernd und Olaf: Vielen Dank für die Einblicke in eure Fachbereiche und die gute Zusammenarbeit.

Meinem Mentoring-Team Christina, Linda und Julia: Vielen Dank fürs Rückenstärken während Tiefs und das gemeinsame Feiern von Hochs.

Meinen FreundInnen: Danke für euer Verständnis, wenn ich Wochenenden am Laptop anstatt mit euch verbracht habe.

Christine und Uli: Euch gilt ein riesengroßes Dankeschön für die Unterstützung in sämtlichen Lebenslagen.

Oma: Danke, dass du schon immer alle meine Pläne bedingungslos unterstützt hast.

Matze: Danke für deine unendliche Geduld und Unterstützung. Du bist der Beste!

Contents

List of Figures

List of Tables

Nomenclature

Acronyms

(a)PCE	(Arbitrary) polynomial chaos expansion
BF	Bayes factor
BMA	Bayesian model averaging
BME	Bayesian model evidence
BMS	Bayesian model selection
CDF	Cumulative distribution function
ES	Energy score
iid	Independent and identically distributed
LOOCV	Leave-one-out cross-validation
MC	Monte Carlo
MJA	Model justifiability analysis
MSA	Model similarity analysis
OM	Original model
PDF	Probability density function

QoI	Quantity of interest
SM	Surrogate model

Greek Symbols

ϵ	Measurement error
μ	Mean
σ	Standard deviation
Ψ	Polynomial of the multivariate orthogonal polynomial basis
Ω_i	Parameter space of model i

Roman Symbols

$\mathbf{c}$	Coefficient vector
d	Maximum polynomial degree of the expansion
D	Number of polynomials considered
$d(x, y)$	distance between x and y
$\mathbb{E}[\cdot]$	Expected value
M	Model
$\tilde{M}$	Surrogate (approximation) of model M
N_m	Number of models in the model set
N_{MC}	Number of Monte Carlo runs
N_p	Number of model parameters
$p(\cdot)$	Prior probability density function

$p(\cdot\|\cdot)$	Posterior probability density function
$P(M_i)$	Prior weight of model M_i
$P(M_i\|\mathbf{y_0})$	Posterior weight of model M_i
$\mathbf{R}$	Covariance Matrix
R^2	Coefficient of determination
$\mathbf{u}$	Model parameters
$Var[\cdot]$	Variance
$\mathbf{x}$	Model inputs
$\mathbf{y}$	Predicted data set
$\mathbf{y_0}$	Reference data set (measured or "synthetic truth")

Abstract

Modeling environmental systems is typically limited by an incomplete system understanding due to scarce and imprecise measurements. This leads to different types of uncertainties, among which conceptual uncertainty plays a key role, but is difficult to address. Conceptual uncertainty refers to the problem of finding the most appropriate model representation of the physical system. This includes the problem of choosing from several plausible model hypotheses, but also the problem that the *true* system description might not even be among this set of hypotheses. In this thesis, I address the first of these issues, the uncertainty of choosing a model from a finite set. To account for this uncertainty of model choice, modelers typically use multi-model methods. This means that they consider not only one but several models and apply statistical methods to either combine them or select the most appropriate one. For any of these methods, it is crucial to know how similar the individual models are. But even though multi-model methods have become increasingly popular, no methods were available that quantify the similarities between models and visualize them intuitively. This dissertation aims at closing these gaps. In particular, it tackles the challenges of judging whether simplified models are a suitable replacement for a more detailed model, and of visualizing model similarities in a way that helps modelers to gain an intuitive understanding of the model set. I defined three research questions that address these challenges and form the basis of this thesis.

1. How can we systematically assess how similar conceptually simplified model versions are compared to an original, more detailed model?

2. How can we extend the similarity analysis so it is suitable for computationally expensive models?

3. How can we visualize the similarities between probabilistic model predictions?

With the first contribution, I show that the so-called model confusion matrix can be used to quantify model similarities and thus identify the best conceptual simplification of a detailed reference model. This matrix was introduced by Schöniger et al. [2015] to estimate the data need of competing models. Here, I demonstrate that the matrix can be used, beyond this original purpose, to analyze model similarities.

With the second contribution, I address the problem of assessing this matrix for computationally expensive models. Since calculating this matrix requires many model runs, the existing method was not yet suitable for models that have long run times. This problem is solved by extending the surrogate-based Bayesian model selection [Mohammadi et al., 2018] so that two models can be compared based on their surrogates while accounting for approximation errors.

With the third contribution, I demonstrate how the similarity of probabilistic model predictions can be quantified based on so-called energy statistics. By comparing different visualization techniques, I show how multi-model ensembles can be visualized intuitively so that modelers can get a better understanding of the model set.

The presented methods are widely applicable and can thus help to bring the importance of model similarities further into the focus of multi-model developers and users. Thus, depending on the research problem, the individual models or an appropriate multi-model method can be selected in a more targeted manner.

Zusammenfassung

Die Modellierung von Umweltsystemen wird in der Regel durch ein unvollständiges Systemverständnis aufgrund von wenigen und ungenauen Messdaten erschwert. Dies führt zu verschiedenen Arten von Unsicherheiten, unter denen die konzeptionelle Unsicherheit zwar eine Schlüsselrolle spielt, jedoch schwer zu berücksichtigen ist. Konzeptionelle Unsicherheit bezeichnet das Problem, die geeignetste Modellrepräsentation des physikalischen Systems zu finden. Dies umfasst zum einen die Wahl zwischen mehreren plausiblen Modellhypothesen und zum anderen das Problem, dass die *wahre* Systembeschreibung möglicherweise nicht unter diesen Hypothesen ist. In dieser Arbeit befasse ich mich mit dem ersten dieser Probleme, der Unsicherheit bei der Wahl eines Modells aus einer endlichen Menge. Um dieser Unsicherheit Rechnung zu tragen, werden häufig Multi-Modell-Methoden verwendet. Das bedeutet, dass man nicht nur ein, sondern mehrere Modelle berücksichtigt und statistische Methoden anwendet, um diese entweder zu kombinieren oder das geeignetste auszuwählen. Für jede dieser Methoden ist es entscheidend wie ähnlich die einzelnen Modelle untereinander sind. Trotz der zunehmenden Anwendung von Multi-Modell-Methoden gab es bisher jedoch keine Methode, die die Ähnlichkeiten zwischen den Modellen quantifiziert und intuitiv visualisiert. Diese Dissertation hat zum Ziel diese Lücke zu schließen. Sie befasst sich insbesondere mit den Fragen, ob vereinfachte Modelle ein geeigneter Ersatz für ein detaillierteres Modell sind, und wie wir Modellähnlichkeiten so visualisieren können, dass Modelliererinnen und Modellierer ein intuitives Verständnis des verwendeten Modellensembles erlangen. Ich habe drei Forschungsfragen definiert, die sich mit diesen Herausforderungen befassen und die Grundlage dieser Arbeit bilden:

1. Wie können wir systematisch erfassen, wie ähnlich konzeptionell vereinfachte Modelle im Vergleich zu einem detaillierten Referenzmodell sind?

2. Wie können wir die Modellähnlichkeitsanalyse so erweitern, dass sie für rechenzeitintensive Modelle geeignet ist?

3. Wie können wir die Ähnlichkeiten zwischen probabilistischen Modellvorhersagen visualisieren?

Mit dem ersten Beitrag zeige ich, dass die sogenannte Modell-Konfusionsmatrix zur Quantifizierung von Modellähnlichkeiten und damit zur Identifikation der besten Vereinfachung eines detaillierten Referenzmodells verwendet werden kann. Diese Matrix wurde von Schöniger et al. [2015] verwendet, um den Datenbedarf konkurrierender Modelle abzuschätzen. In der vorliegenden Arbeit zeige ich, dass die Matrix über diesen ursprünglichen Zweck hinaus zur Analyse von Modellähnlichkeiten genutzt werden kann.

Mit dem zweiten Beitrag gehe ich auf das Problem der Berechnung dieser Matrix für rechenzeitintensive Modelle ein. Da die Erstellung der Matrix viele Modellläufe erfordert, war diese Methode für rechenzeitintensive Modelle nicht geeignet. Dieses Problem wird durch die Erweiterung einer surrogatbasierten Bayes'schen Analyse [Mohammadi et al., 2018] gelöst, sodass zwei Modelle basierend auf ihren Surrogaten unter Berücksichtigung von Approximationsfehlern verglichen werden können.

Mit dem dritten Beitrag zeige ich, wie die Ähnlichkeit probabilistischer Modellvorhersagen auf Grundlage der sogenannten „energy statistics“ quantifiziert werden kann. Durch den Vergleich verschiedener Visualisierungstechniken stelle ich außerdem dar, wie Modellensembles intuitiv visualisiert werden können, um ein besseres Verständnis des gewählten Modellraums zu ermöglichen.

Die vorgestellten Methoden sind vielseitig anwendbar und können so helfen, die Bedeutung von Modellähnlichkeiten für Multi-Modell-Methoden weiter in den Fokus von Entwicklerinnen und Anwendern zu bringen. Dadurch können, je nach Problemstellung, die einzelnen Modelle oder eine geeignete Multi-Modell-Methode gezielter ausgewählt werden.

1 Introduction

Modelers in geoscience face two sources of uncertainties: inherent randomness and lack of knowledge [e.g. Ayyub and Klir, 2006, Rinderknecht et al., 2012].

- **Inherent randomness**: The system is truly random, i.e. its behavior is not predictable and the uncertainty cannot be reduced by more observations. This type of uncertainty is also known as the aleatory component of the total uncertainty [e.g. Ayyub and Klir, 2006].
- **Lack of knowledge**: The system is not fully understood because of scarce and imprecise data that does not sufficiently reflect the heterogeneity in space, variability in time, and interacting processes that might occur on different scales. This leads to uncertainty in the time-varying input variables, in the parameters (time-invariant input variables), and in the model concept. This type of uncertainty is referred to as epistemic or hypothetical uncertainty [e.g. Ayyub and Klir, 2006, Nearing and Gupta, 2018]. Theoretically, it can be overcome by collecting more (informative) data. However, this might not always be possible in practical applications due to limited resources or technical means. In geoscience and engineering, this type of uncertainty is usually the dominant one [Rinderknecht et al., 2012, Ayyub and Klir, 2006].

In this work, I focus on the uncertainty in choosing the most appropriate representation of the real system (conceptual uncertainty), while accounting for the effects of limited and imprecise data. The following section specifies conceptual uncertainty in general and its scope within this thesis. Subsequently, I introduce typical approaches that deal with conceptual uncertainty and their associated issues.

1.1 Conceptual Uncertainty

To define conceptual uncertainty, it is helpful to recap the different stages of model building. The process of model development can be divided into three main stages according to Gupta et al. [2012]:

1. **The conceptual model**:
   This first stage includes the definition of the system boundaries, the relevant variables, the time-invariant properties of the system and their spatial variability, conservation laws, underlying assumptions as well as uncertainties concerning each of these components [Gupta et al., 2012]. Please refer to Enemark et al. [2019] for a detailed example of the components of the conceptual model in the context of hydrogeology.

2. **The mathematical model**:
   The second stage incorporates the equations that describe the dynamics of the variables and their interactions. It also comprises the representation of spatial variability [Gupta et al., 2012].

3. **The computational model**:
   In the third stage, the mathematical equations are transferred to a numerical description. For distributed, dynamic systems, the modeler has to choose a spatial discretization method, the resolution as well as a method for the time integration [Gupta et al., 2012].

While input uncertainties and parameter uncertainty can be quantified by propagating probability density functions (PDFs) through the model to the prediction of the quantities of interest, this is impossible for conceptual uncertainty [Abramowitz, 2010]. Instead, multi-model approaches are commonly used to address the uncertainty of model choice. However, it is important to note that there is no multi-model method that can quantify conceptual uncertainty on an absolute level [Nearing and Gupta, 2018]. The reasons for this are evident: In practice, there is no way to create and sample from an exhaustive list of all plausible models that covers the entire range of possible outcomes [Vehtari and Ojanen, 2012, Ferré, 2017, Nearing and Gupta, 2018, Höge et al., 2019]. Therefore, Nearing and Gupta [2018] suggest understanding multi-model

methods rather as sensitivity analyses. From this point of view, we can recognize multi-model methods as tools that make modelers aware of how much their results may differ within the discrete, finite set of models under consideration. This follows the line of thought of Ferré [2017] who introduced the idea of a multi-model ensemble as a "team of rivals", which provides competing views of a system. If the competing models agree on a certain prediction, this increases the decision makers' confidence, while disagreement indicates the need for further investigation [Ferré, 2017]. This leads to the challenge of building such a "team of rivals", which is a model set that covers a broad range of plausible system behavior.

Refsgaard et al. [2012] propose to use the so-called "MECE criterion" for selecting a model set. According to this criterion, models should be mutually exclusive and collectively exhaustive. The first requirement, mutually exclusive, means that the models should be based on independent hypotheses. This corresponds to the "team of rivals" idea of Ferré [2017]. As model independence is a key concept for multi-model ensembles, it will be discussed in more detail in Section 1.3. The second requirement, i.e. being collectively exhaustive, corresponds to a model set that covers the full plausible range of possible system behavior. This means that such a set would include even the "unknown unknowns" [Enemark et al., 2019], which are hypotheses describing potential system behavior that has not yet been observed. For this reason, this criterion cannot be fulfilled in practice [Refsgaard et al., 2012]. Therefore, the methods described in this thesis deal with the uncertainty of choosing from a discrete, finite set of models.

Enemark et al. [2019] criticize that studies on multi-model methods often lack a systematic approach for the development of the individual models. The authors recap three methods that guide the model development:

1. **Varying complexity**: all models in the set are based on the same concept and the complexity is gradually increased or decreased. An appropriate level of complexity can be found using a so-called model justifiability analysis as presented in Schöniger et al. [2015].

2. **Alternative interpretation**: different models are built by different experts based on the same measurement data. This strategy, however, can result in very similar or very different models and there is no systematic way to control the diversity of the models.

3. **Hypothesis testing**: the models are built such that the difference between their conceptualizations is maximized, i.e. a so-called antithetic setup is defined.

Enemark et al. [2019] emphasize that only the hypothesis testing method attempts to meet the mutually exclusive criterion. However, it is not guaranteed that it is fulfilled, and for real applications, it can be hard to define models as antithetic hypotheses.

The varying complexity approach does not seek to define models in a way that they are mutually exclusive. It rather checks whether the models' complexities are in balance with a realistic amount of available calibration data. This is necessary because a more complex model usually needs more (informative) data to constrain its parameters during its calibration [e.g. Guthke, 2017, Höge et al., 2018]. Therefore, model complexity and effort for acquiring field data have to be balanced. If the model is too complex for a given amount of calibration data, it will show a high variance in its predictions, this effect is known as overfitting [e.g. Babu, 2011, Lever et al., 2016, Höge et al., 2019]. In the opposite case, a too simple model, which needs less data for calibration, will show a high bias in its predictions and thus underfits the system [e.g. Babu, 2011, Lever et al., 2016, Höge et al., 2019]. This issue is well-known as "bias-variance-tradeoff" [e.g. Geman et al., 1992, Burnham and Anderson, 2002]. Consequently, for a certain amount of measurement data, there is a level of model complexity that is already complex enough to capture all relevant processes (if visible in the data) but still simple enough so that it does not overfit ("principle of parsimony") [Schöniger et al., 2015].

1.2 Multi-Model Methods

Once the model set is defined, an appropriate multi-model technique has to be chosen. Höge et al. [2019] highlight the need to clearly specify the objective of the applied multi-model method because the results may vary considerably between different objectives and their corresponding methods.

There are numerous multi-model approaches. Each one of them is tailored to a particular purpose in modeling and is founded on different philosophical perspectives. Höge et al. [2019] provide guidelines for choosing the right method for a specific multi-model problem. Here, I briefly recap Bayesian model averaging, Bayesian model selection as well as Bayesian model combination and their scope of application:

Bayesian model averaging (BMA) [Hoeting et al., 1999] and Bayesian model selection (BMS) [Raftery, 1995] are well-known approaches to address the uncertainty of model choice. Both approaches are based on the same mathematical framework but they differ in their purpose [Höge et al., 2019]. For both methods, it is assumed that the "true" (data-generating) model is contained in the model set (the collectively exhaustive assumption) and model weights are calculated that reflect the probability of each model to be the true one. In the Bayesian model averaging framework, the model weights are used to build a weighted average of the individual model predictions [Wöhling et al., 2015]. It is important to note that building this weighted average is not a model combination method that aims at finding the combination of several models that fits the measured data best [Minka, 2002, Monteith et al., 2011, Höge et al., 2019]. BMA should rather be understood as a method that quantifies the uncertainty of finding the true model within the set as long as the data set size does not suffice for clear identification. In the limit of infinite data set size, Bayesian model selection (BMS) will identify this true model [Höge et al., 2019] by assigning it a weight of one. In the case of finite measurement data, however, the identification of the true model may be impossible because two or more models receive similar weights. Therefore, Minka [2002] calls BMA an "intermediate step on the way to BMS".

Bayesian model combination (BMC) takes an entirely different perspective: If our primary goal is to make robust predictions of the system behavior and we do not assume that our set contains the true model, we should consider combining models. In contrast to model averaging, which traditionally acts on PDFs, model combination acts on the level of pointwise forecasts [Monteith et al., 2011]. There are different model combination methods available and detailed comparisons can be found in Diks and Vrugt [2010], Arsenault et al. [2015] and Höge et al. [2019].

All of the mentioned techniques have been applied numerously in different fields of science and engineering. Applications in geoscience include, among others, modeling climate [Tebaldi and Knutti, 2007, Abramowitz and Bishop, 2015], weather [Krishnamurti et al., 2000], soil-plant-atmosphere systems [Wöhling et al., 2015, Wallach et al., 2018], groundwater [Rojas et al., 2008, Schöniger et al., 2015], reactive transport in groundwater [Lu et al., 2015], vadose zone hydrology [Wöhling and Vrugt, 2008, Diks and Vrugt, 2010] and streamflow hydrology [Georgakakos et al., 2004, Ajami et al., 2007, Diks and Vrugt, 2010, Arsenault et al., 2015, Schöniger et al., 2014].

For each of these methods, it is crucial to know how (dis-)similar the individual models are. In the following sections, I want to clarify what is meant by model similarity or dependence (both terms are used in the literature) and why it is important for multi-model approaches.

1.3 Model Similarity

Types of Model Similarity

Two models can be described as similar in their structure or their predictions.

Structural similarity means that models share parts of their conceptualization because they are based on similar hypotheses. This type of similarity contradicts the "mutually exclusive" criterion and undermines the "collectively exhaustive" criterion (see Section 1.1).

Prediction similarity implies that the models produce similar predictions when similar inputs are used. This can be caused by structural similarity of the models but can also happen by chance.

In literature, the term "model dependence" is often used synonymously with "model similarity" [e.g. Abramowitz and Gupta, 2008, Bishop and Abramowitz, 2013, Abramowitz and Bishop, 2015, Sanderson et al., 2015a, Knutti et al., 2017]. In this context, dependence is not used in a strict mathematical sense but rather means that two models are considered dependent if they are (partly) based on common hypotheses. In this thesis, I subsume "model dependence" under *structural similarity* to distinguish it from *prediction similarity*.

In the following chapters, I present methods to quantify *prediction similarity* based on probabilistic model outputs because it allows a direct comparison of different models and of the models and measurement data.

The Role of Model Similarity in Model Development

Similarity is important during all stages of the modeling process from the composition of a model set and the choice of an appropriate multi-model method to the definition and interpretation of model weights.

In the first step, most multi-model methods require assigning weights to the individual models. In the framework of Bayesian model selection or averaging, modelers have to define prior weights that represent their belief in each model's adequacy. In a second step, these prior weights are updated to posterior weights based on the model's ability to reproduce a calibration data set. In the following, I describe why model similarity is highly important for the definition of the prior weights and for the interpretation of the posterior weights.

Choosing Prior Weights for Similar Models

A common choice for prior weights is the uniform distribution, which means that all models are assumed to be equally plausible. However, this leads to difficulties when there are models in the set that are very similar: If we imagine the extreme case of adding a model to the set that is identical to one already contained, giving equal weights to each model would double the weight for the duplicate models. George [2010] and Garthwaite and Mubwandarikwa [2010] therefore recommend to use so-called "dilution priors" that divide the weight between partly redundant models. Both studies suggest different methods for constructing dilution priors, some of them limited to linear models. They also mention a method based on distances between models that is more general and can also be applied to non-linear models. As one possible approach to calculate the distance between two models, George [2010] proposes to use the Hellinger distance [Hellinger, 1909] based on the marginal distributions of the data given a model. However, the author does not provide details about the implementation and the choice of the metric. With the work presented here, I want to further pursue the idea of quantifying the similarity between two models by measuring the distance between them. Hence, in future work, the presented methods can be useful for defining dilution priors.

Uncovering Similarity from Posterior Weights

Updating the prior weights with observational data yields posterior weights. Modelers

might run into difficulties when they try to interpret these posterior weights: Two models might receive similar posterior weights either because their predictions are actually similar or because one model fits the observations better but has a higher variability compared to its competitor. The reason for the latter case is that the posterior model weights reward models for high goodness-of-fit and punish them for high flexibility [Schöniger et al., 2015]. Each of these situations requires a different multi-model approach. Therefore, a similarity analysis is helpful to uncover which scenario has led to similar posterior weights.

Effect of Similarity on the Performance of Multi-Model Ensembles

Aside from the definition of prior weights or the interpretation of posterior weights, many authors have assessed model similarity in the context of multi-model ensemble performance. They compared the predictive performance of multi-model ensembles to the performances of the individual ensemble members. It has been highlighted that the superiority of the multi-model ensemble depends strongly on the dissimilarity of the individual models [Tebaldi and Knutti, 2007, Abramowitz and Gupta, 2008, Abramowitz, 2010, Winter and Nychka, 2010, Bishop and Abramowitz, 2013, Sanderson et al., 2015a, Abramowitz et al., 2018, Enemark et al., 2019]. Other important factors for the multi-model ensemble performance are the applied combination/averaging methods [Krishnamurti et al., 2000, Doblas-Reyes et al., 2005, Diks and Vrugt, 2010, Arsenault et al., 2015, Höge et al., 2019, 2020], the properties of the individual models such as prediction spread [Fritsch, 2000, Doblas-Reyes et al., 2005, Weigel et al., 2008] and the performance measure [Hagedorn et al., 2005, Weigel et al., 2008, Diks and Vrugt, 2010].

A question one often comes across in the context of multi-model ensembles is whether the ensemble performs better than its best individual member. However, as Hagedorn et al. [2005] point out, before asking whether the ensemble outperforms the individual models, one should ask whether a "best" individual model even exists. Usually, this is not the case, as each model has its individual strengths and weaknesses and none of them performs best under all relevant scenarios and considering all quantities of interest [Hagedorn et al., 2005]. Therefore, the authors conclude that the strength of a multi-model ensemble is its robustness, i.e. it performs well over a broad range of output variables and boundary conditions.

Therefore, for making general statements on the performance of single models or multi-

model ensembles, I suggest evaluating each model's *overall* skills considering all relevant output variables at all important points in the spatial and temporal domain. The methods presented in this thesis enable modelers to objectively compare the individual models and the ensemble (possibly generated by different model combination methods) and to identify the one that is closest to the observed data.

1.4 Structure of This Thesis

This thesis is a cumulative dissertation based on three journal articles/manuscripts. After the preceding introduction in Chapter 1, I give an overview of the state of the art in Chapter 2. Objectives and contributions of this thesis are outlined in Chapter 3, results are summarized in Chapter 4. The conclusions drawn from these results and an outlook to possible future research are given in Chapter 5. Appendix A contains the three articles/manuscripts that were prepared in the course of this project.

2 State of the Art

This chapter summarizes the state of the art that served as a starting point for my contributions. In section 2.1, the mathematical framework of Bayesian model selection (BMS) and the so-called model justifiability analysis are presented, which form the basis for subsequent analysis tools. This is followed by an extension of BMS for computationally expensive models in Section 2.2. Section 2.3 describes methods for quantifying and visualizing model similarities. This includes a general introduction to probability metrics, followed by a presentation of specific tools (energy distance and energy score) and an overview of commonly used visualization tools that will be adapted for an intuitive representation of model similarities.

2.1 Ranking Models Based on BME

2.1.1 Mathematical Framework of BMS

Bayesian model selection [Hoeting et al., 1999] is a commonly used method to identify the most appropriate model from a set of competing model alternatives. BMS ranks models based on weights, which denote the probability of model M_k to have generated the data $\mathbf{y}_0$ [e.g. Raftery, 1995, Wasserman, 2000].

First, each model in the set obtains a prior model weight $P(M_k)$. In Bayesian statistics, a prior probability reflects the modeler's belief based on expert knowledge. It is formulated before measurement data $\mathbf{y}_0$ are taken into account. In the BMS framework, a typical choice are uniform prior weights $P(M_k) = 1/N_m$ (with N_m being the number of alternative models) that treat all models in the set as equally likely. However, this choice is questionable if two or more models are very similar. In this case, uniform priors would distort the weights of these models. To address this problem, Garthwaite

and Mubwandarikwa [2010] and George [2010] suggest to use so-called dilution priors that account for redundancies in the model set.

The prior weights $P(M_k)$ are updated to posterior weights $P(M_k|\mathbf{y}_0)$ based on Bayes' theorem:

$$P(M_k|\mathbf{y}_0) = \frac{p(\mathbf{y}_0|M_k)\,P(M_k)}{\sum_{i=1}^{N_m} p(\mathbf{y}_0|M_i)\,P(M_i)}, \tag{2.1}$$

in which $p(\mathbf{y}_0|M_k)$ is the so-called Bayesian model evidence (BME). BME is well suited for ranking models because it implicitly performs a trade-off between goodness-of-fit and model parsimony [Schöniger et al., 2014]. BME is also known as marginal likelihood because it can be calculated by averaging (marginalizing) over the model's parameter space $\mathcal{U}_k$ [Kass and Raftery, 1995, Schöniger et al., 2014]:

$$p(\mathbf{y}_0|M_k) = \int_{\mathcal{U}_k} p(\mathbf{y}_0|M_k, \mathbf{u}_k)\, p(\mathbf{u}_k|M_k)\, d\mathbf{u}_k, \tag{2.2}$$

where $\mathbf{u}_k \in \mathcal{U}_k$ is the parameter vector of model M_k, $p(\mathbf{u}_k|M_k)$ is the prior parameter distribution, and $p(\mathbf{y}_0|M_k, \mathbf{u}_k)$ is the likelihood, i.e. the probability of the model M_k with parameter set $\mathbf{u}_k$ to have generated the data set $\mathbf{y}_0$.

BME thus quantifies the model's likelihood to have generated the data $\mathbf{y}_0$ independent of the parameter choice. A straight-forward way to approximate Equation 2.2 is to sample the prior distribution of the model parameters $\mathbf{u}_k$ using brute-force Monte Carlo and to evaluate the likelihood $p(\mathbf{y}_0|M_k, \mathbf{u}_k)$ of the measured data $\mathbf{y}_0$ for each parameter vector $\mathbf{u}_{k,i}$ [Schöniger et al., 2014]:

$$p(\mathbf{y}_0|M_k) \approx \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} p(\mathbf{y}_0|M_k, \mathbf{u}_{ki})\,. \tag{2.3}$$

2.1.2 Bayes Factor

Apart from ranking models based on Bayesian model weights, as it is done in BMS, the decisiveness of the choice between two competing models can be expressed using so-called Bayes factors [Jeffreys, 1961, Kass and Raftery, 1995]. The Bayes factor between two models is defined as the ratio of their respective BME values:

$$BF\left(M_i, M_j\right) = \frac{P\left(M_i|\mathbf{y}_0\right)}{P\left(M_j|\mathbf{y}_0\right)} \frac{P\left(M_j\right)}{P\left(M_i\right)} = \frac{p\left(\mathbf{y}_0|M_i\right)}{p\left(\mathbf{y}_0|M_j\right)}. \tag{2.4}$$

Jeffreys [1961] introduced categories for interpreting the Bayes factor as evidence against M_j. I will use the slightly modified scale suggested by Kass and Raftery [1995] as shown in Table 2.1.

Table 2.1: 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

Accordingly, negative $log_{10}(BF)$ values favor M_j over M_i.

2.1.3 Model Justifiability Analysis

Another way to analyze models based on BMS is the so-called model justifiability analysis (MJA) that was introduced by Schöniger et al. [2015]. The method determines the maximum level of model complexity that is affordable given a realistic amount of measurements before these data have been actually measured.

The analysis is based on a so-called "model confusion matrix" (see Figure 2.1). 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 predicted classifications. Thus, it is easily visible whether an object is misclassified ("confused").

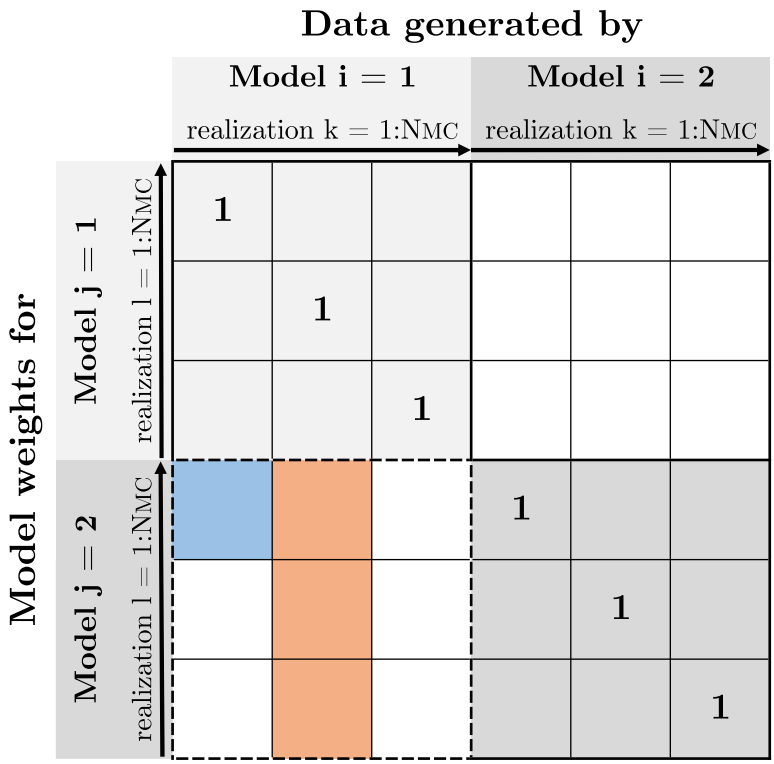


Figure 2.1: Schematic illustration of the model confusion matrix for two models M_1 and M_2, modified after Schäfer Rodrigues Silva et al. [2020]. **Blue box**: likelihood of a single realization drawn from M_2 given a realization $k = 1$ drawn from M_1. **Red box**: BME value (average likelihood) of M_2 given a single realization $k = 2$ of M_1. This BME value is normalized by the sum of the BME values of all models for this data set, which yields a single model weight. **Dashed box**: Averaging these weights over all synthetic data sets $k = 1, ..., N_{MC}$ of the data-generating model M_1 yields the model weight $P(M_2|M_1)$, i.e. the expected weight of M_2 given that M_1 is true.

The same concept is transferred to the problem of model identification. The core idea of the justifiability analysis is that the models are tested against each other instead of testing them against measurements. Each of the models in the set is sequentially treated as "synthetic truth" and the other models are tested against this "true" one. Then it can be tested how well a model can be identified as the data-generating process based on the Bayesian model weights (see Equation (2.1)): If a model receives the highest weights for its own predictions, it is able to self-identify its data. From a successful self-identification, we can conclude that the model's complexity is legitimate given the current experimental setup. In contrast, if a model's predictions are "confused" with those of the other models, its complexity cannot be justified.

Implementation of the Model Confusion Matrix

In a Monte Carlo loop, each of the $i = 1, ..., N_M$ models takes turns to generate $k = 1, ..., N_{MC}$ data sets $\mathbf{y}_{0,ik}$ by drawing samples of their prior parameter distributions $p(\mathbf{u}_i|M_i)$. These data sets are treated as "synthetic truth". In this role, I refer to the models as "data-generating" and list them in the columns of the model confusion matrix (see Figure 2.1). The same data sets are listed in the rows of the matrix. Here, however, the models' roles are different: They are not considered as data-generating processes but as models to be evaluated against the synthetic truth.

This evaluation is done by calculating 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 evaluated model M_j with the parameter vector $\mathbf{u}_{jl}$ (blue box in Figure 2.1).

Averaging these likelihoods over all parameter realizations $l = 1, ..., N_{MC}$ (rows) of the evaluated model M_j yields a BME value (see Equation 2.3, red box in Figure 2.1): $\text{BME}_{j,ik} = p(\mathbf{y}_{0,ik}|M_j)$.

This BME value is normalized by the sum of the BME values of all models given this reference data set $y_{0,ik}$ (see Equation 2.1). The normalization yields a single posterior model weight $P\left(M_j|\mathbf{y}_{0,ik}\right)$.

This weight depends on the reference data set $y_{0,ik}$ (column) sampled from the data-generating model M_i. Therefore, we also average over $k = 1, ..., N_{MC}$ columns of the data-generating model M_i to get the mean posterior weight of model M_j given the data produced by model M_i:

$$P(M_j|M_i) = \frac{1}{N_{MC}} \sum_{k=1}^{N_{MC}} \frac{\text{BME}_{j,ik}}{\sum_{j=1}^{N_m} \text{BME}_{j,ik}}. \tag{2.5}$$

This Bayesian model weight quantifies the probability of the model M_j to be the most appropriate one from the set if M_i was true. It corresponds to one entry of the final model confusion matrix (dashed box in Figure 2.1). The matrix is symmetric with size $N_M \times N_M$. Its main-diagonal entries are the self-identification weights. The off-diagonal entries can be interpreted as a measure of similarity between two models. For an infinite data set size, the respective data-generating model will be identified with a weight of

100% (unless two models would yield exactly identical predictive distributions). For finite data sets, we can observe that, for increasing model complexity, more and more data is needed for a self-identification [Schöniger et al., 2015].

2.2 BMS for Computationally Expensive Models

2.2.1 Surrogate Modeling

The standard evaluation of BME using a brute-force Monte Carlo sampling of the parameter prior distributions involves a large number of model evaluations. For computationally expensive models, this soon becomes intractable. To address this problem, so-called surrogate models can be used, which approximate the original model and have reduced run times. A surrogate model maps the outputs of the original model given its parameters. It can be seen as a black-box model because it does not involve any knowledge of the physical system.

There are different types of surrogate models. Amongst the most commonly used methods are Gaussian process emulators (also known as "Kriging" in the context of geostatistics) [Krige, 1951, Matheron, 1963, Sacks et al., 1989], polynomial chaos expansion (PCE) [Wiener, 1938, Xiu, 2010], radial basis functions [Buhmann, 2003], support vector machines [Vapnik, 2000] and neural networks [Bishop, 1995].

In a recent benchmark study, Köppel et al. [2019] compared five data-driven methods for building surrogate models regarding their use for uncertainty quantification in geosciences, their computational costs to construct and evaluate the surrogate, their accuracy for a low or high number of model runs, and their applicability to high-dimensional problems. Their study revealed that the arbitrary polynomial chaos expansion (aPCE) method is very commonly used in geoscience, has comparatively low computational costs for constructing and evaluating the surrogate, and has good accuracy for a low number of model runs. Therefore, it is well suited for the application in this study.

2.2.2 aPCE-Based Surrogate Modeling

The data-driven aPCE approximates the model output by its dependence on model parameters via multivariate polynomials. In classical PCE, polynomial families are chosen according to the distribution of the parameters (e.g. uniform or Gaussian) [e.g. Marelli and Sudret, 2021]. However, in real-world applications, it is a commonly known problem that variables cannot be described by one of these probability distributions or only limited data is available, so it is impossible to fit a unique parametric probability distribution [Oladyshkin and Nowak, 2012]. aPCE overcomes these shortcomings as it requires no approximation of a density function.

As the aPCE-based surrogates used for this study were built using existing code by Oladyshkin [2020], their construction is not within the scope of this thesis. Therefore, I will only give a brief overview of the underlying theory in the next section. For implementation details, I refer to publication P2 [Scheurer et al., 2021] in appendix A.2.

Based on the original polynomial chaos expansion introduced by Wiener [1938], the aPCE constructs surrogate models with the help of an orthonormal polynomial basis:

$$M(\boldsymbol{x}, t; \mathbf{u}) \approx \tilde{M}(\boldsymbol{x}, t; \mathbf{u}) = \sum_{s=1}^{D} c_s(\boldsymbol{x}, t) \cdot \Psi_s(\mathbf{u}), \tag{2.6}$$

where $\tilde{M}(\boldsymbol{x}, t; \mathbf{u})$ is the surrogate that approximates the original model's outputs $M(\boldsymbol{x}, t; \mathbf{u})$ and $\Psi_s(\mathbf{u})$ represents the multivariate polynomial basis. The coefficients $c_s(\boldsymbol{x}, t)$ are the corresponding coordinates, i.e. they quantify how the model response depends on the parameters $\mathbf{u}$ for each point in space and time [Köppel et al., 2019]. In standard settings, the number of polynomials D depends on the number of model parameters N_p and the chosen maximum polynomial degree d according to $D = (Np + d)!/(Np!d!)$.

The polynomials are constructed according to the method described by Oladyshkin and Nowak [2012]. To compute the coefficients $c_s(\boldsymbol{x}, t)$, a non-intrusive stochastic collocation method [Oladyshkin et al., 2012] is used. 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.

The procedure described so far yields a surrogate based on the prior distribution of the parameters and does not include information from measurements. Therefore, the constructed surrogate model $\tilde{M}$ might produce imprecise responses for regions in 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, an iterative Bayesian updating of the aPCE representation (BaPCE) is used. It improves the accuracy of the surrogate by incorporating new collocation points at approximate locations of the maximum posterior parameter set [Oladyshkin et al., 2013].

2.2.3 aPCE-Based BMS

Mohammadi et al. [2018] first used aPCE-based surrogate models to enable BMS for computationally expensive models. To account for the approximation error that is introduced when the original model is replaced by a surrogate model, Mohammadi et al. [2018] introduced the corrected BME:

$$p(\boldsymbol{y}_0|M_k) = p(\boldsymbol{y}_0|\tilde{M}_k) \cdot \int_{\mathcal{U}_k} p(M_k|\tilde{M}_k, \mathbf{u})\ p(\mathbf{u}|\tilde{M}_k, \boldsymbol{y}_0)\ d\mathbf{u}. \tag{2.7}$$

Equation (2.7) shows how the BME value of the original model ($\mathrm{BME_{OM}}$) can be calculated from the BME value of the surrogate model ($\mathrm{BME_{SM}}$) and a correction factor:

$$\mathrm{BME_{OM}} = \mathrm{BME_{SM}} \cdot \mathrm{Weight_{SM}}, \tag{2.8}$$

with

$$\begin{aligned} \mathrm{BME}_{\mathrm{OM}} &= p(\boldsymbol{y}_0|M_k), \\ \mathrm{BME}_{\mathrm{SM}} &= p(\boldsymbol{y}_0|\tilde{M}_k) \text{ and} \\ \mathrm{Weight}_{\mathrm{SM}} &= \int_{\mathcal{U}_k} p(M_k|\tilde{M}_k, \mathbf{u})\ p(\mathbf{u}|\tilde{M}_k, \boldsymbol{y}_0)\ d\mathbf{u}, \end{aligned} \tag{2.9}$$

where the $\mathrm{BME}_{\mathrm{SM}}$ value can be computed as described in Section 2.1, using the surrogate model $\tilde{M}_k$ instead of the original model M_k. The correction factor $\mathrm{Weight}_{\mathrm{SM}}$ requires an integration over the posterior parameter space $\mathcal{U}_k$. As the original models are usually computationally too expensive for using Monte Carlo integration, the correction factor can be estimated at P collocation points $\mathbf{u}^*$ that were used to construct the surrogate model [Mohammadi et al., 2018]:

$$\mathrm{Weight}_{\mathrm{SM}} \approx \sum_{i=1}^{P} p(M_k|\tilde{M}_k, \mathbf{u}_i^*)\ p(\mathbf{u}_i^*|\tilde{M}_k, \boldsymbol{y}_0). \tag{2.10}$$

2.3 Quantifying and Visualizing Model Distances

As outlined in Section 1.3, model similarity plays a crucial role for multi-model methods. Though it has been widely discussed (see Section 1.3 and references therein), only a few publications suggested methods to quantify and visualize model similarities. In this section, I will give an overview of existing approaches, followed by the mathematical basics of the method I present in publication P3 (see appendix A.3).

2.3.1 Definitions of Model Spaces

In the context of model similarities, one often comes across the term "model space" [e.g. Abramowitz and Gupta, 2008, Abramowitz, 2010, Knutti et al., 2010, George, 2010,

Vehtari and Ojanen, 2012, Höge, 2019]. However, there is no comprehensive definition [Höge, 2019] and hence, it is not used in a consistent way.

Many times, "model space" is used synonymously with model set $\mathcal{M}$, i.e. a finite set of models that build a multi-model ensemble. In other studies, it refers to the set of models and combinations thereof ("interpolating" between the individual models). What remains unclear, however, is how dimensions of such a space could be defined. Höge [2019] suggested considering model properties such as "number of model parameters, degree of non-linearity of functional relations in the model". An attempt to construct such a model space has been presented in Getz et al. [2018], where models are located in a "structure-/process-/utility-complexity space". However, this is a qualitative ranking of the models along the three axes.

In some cases, what is called "model space" might be better described by the term "prediction space", i.e. the space containing the model predictions with the number of quantities of interest being the number of dimensions and the units being the ones of each quantity of interest. When model predictions and measurements are compared, also the term "quantity of interest (QoI) space" can be appropriate as it indicates that this space includes model predictions and observed data. In this space, we can assess how models relate to observations and gain a better understanding of the three "perspectives on model comparison" defined by Bernardo et al. [1999]. These categories classify how the model set $\mathcal{M}$ relates to the observations: $\mathcal{M}$-closed (one of the models in the set is the data-generating process), $\mathcal{M}$-completed (the data-generating process is not in the model set, but can be approximated), and $\mathcal{M}$-open (the data-generating process is not in the model set and cannot even be described given the current background information).

This thesis will analyze model sets through their predictive distributions, similarities and distances in application-specific QoI spaces, as inherent in the Bayesian approach. More details will be provided in Section 2.3.3.

2.3.2 Existing Model Similarity Measures

In the following section, I will briefly outline four studies that introduced "model independence metrics" or "model space distances". However, none of these studies verified

that the suggested measures fulfill metric properties (see Section 2.3.3) [Abramowitz et al., 2018].

Abramowitz [2010] suggested examining the model space by using "projections" of it onto various performance measures. As model performance measures act on the model predictions, I assign their definition rather to the "quantity of interest" category than to an actual "model space". They provided an example of such a projection by introducing the so-called "conditional bias space", in which distances between models can be defined. In their method, the dimensionality of the problem was reduced by clustering the model input and thus defining nine typical conditions. For all time steps that belong to these input clusters, selected output variables were considered. For these output variables, univariate PDFs were fitted and the intersections of these PDFs for pairs of models were calculated, yielding a distance measure. Based on the occurrence of the respective cluster, these distances were weighted and used as a proxy for model similarities.

Sanderson et al. [2015a] used an empirical orthogonal function analysis to reduce the dimensionality of their quantities of interest. This method generates a relatively small number of statistically independent variables from the original high-dimensional model predictions. In the resulting multivariate empirical orthogonal function basis, Euclidean distances between models and observations were calculated and multi-dimensional scaling (MDS) was used to project these distances onto two dimensions. In the resulting low-dimensional representation, it is possible to interpolate between models accounting for model similarity and accuracy [Sanderson et al., 2015a]. MDS [Kruskal, 1964] is commonly used to address dimensionality reduction problems. This method finds a low (often two) dimensional configuration of the objects that matches their original, high-dimensional configuration as well as possible. However, the true distances cannot be perfectly preserved and, consequently, the low-dimensional representation can be misleading.

Knutti et al. [2017] and Lorenz et al. [2018] calculated the root-mean-square error (RMSE) between predicted variables to quantify model similarity and goodness-of-fit to observations. These similarity measures were, however, not used for visualizing the model space, but for defining model weights for a multi-model ensemble.

Unlike the existing methods described so far, Bennett et al. [2019] compared models not asking how much their predictions differ, but why. Using transfer entropy, which is

an information-theoretic measure, the authors were able to quantify how much information is transferred from one variable to another through the model structure. This method can account for non-linear processes and system feedbacks [Bennett et al., 2019]. The results are visualized using chord diagrams that illustrate the information "flow" between variables and, hence, intuitively show structural differences between models.

This is the only method I know that addresses structural similarity. However, testing this method with the data I have used in study P3 (see appendix A.3) revealed that it requires very long time series. Consequently, it seems to be only applicable when extensive data sets are available.

2.3.3 Comparing Probabilistic Model Predictions

In publication P3 (see appendix A.3), I have compared models based on their predictive distributions using probability metrics. The reasons for choosing the methods are the following: (1) Compared to the information theory-based structural comparison of models as introduced by Bennett et al. [2019], fewer data are required to achieve reliable results. (2) Comparing models based on their predictions allows a direct assessment of the goodness-of-fit to observations in the same "quantity of interest space". (3) By using probability metrics, one can account for the predictive uncertainty of the models and measurement errors.

In the following section, I describe how the similarity between two models or between a model and observed data can be quantified and visualized based on samples of the models' predictive distributions. I start with a general introduction to probability metrics, followed by details on the so-called energy distance and energy score, which I have used for comparing models.

Probability Metrics

A function $d : X \times X \to \mathbb{R}$ is called a metric on X if, for all $x, y, z \in X$, it holds:

$$\begin{aligned}
&1.\ d(x,y) \geq 0 \text{ (non-negativity)} \\
&2.\ d(x,y) = 0 \text{ if and only if } x = y \text{ (identity of indiscernibles)} \\
&3.\ d(x,y) = d(y,x) \text{ (symmetry)} \\
&4.\ d(x,y) \leq d(x,z) + d(z,y) \text{ (triangle inequality)}
\end{aligned} \tag{2.11}$$

[e.g. Deza and Deza, 2016].

These axioms can be relaxed in various ways, leading to different generalizations. The definitions of these generalizations are not unique and might differ between different sources. I use the definitions by Deza and Deza [2016]:

A function $d : X \times X \to \mathbb{R}$ is called a quasi-metric on X if d is non-negative and $d(x,x) = 0$ for all $x \in X$. This means that $d(x,y) = 0$ can be fulfilled for some values $x \neq y$. Symmetry as well as the triangle inequality do not need to be fulfilled. In statistics, a quasi-metric is called divergence [e.g. Deza and Deza, 2016].

Distance measures like the Euclidean or Manhattan distance (L_P-metrics) are based on the coordinates of points in the Euclidean space. These distances do not take the density of distributions into account. In contrast, statistical distances (also known as probability metrics) include information about probability densities. Probability metrics measure the distance between two statistical objects such as random variables, probability distributions, or data samples [Martos Venturini, 2015, Deza and Deza, 2016]. The measures that are most commonly used in statistics are the Bhattacharya-, Hellinger-, Kolmogorov-Smirnov (KS) - distance, Kullback-Leibler (KL)- , Jeffrey- (symmetric KL) and χ^2- divergence [Deza and Deza, 2016]. Typical applications of distances in statistics and data analysis are hypothesis tests, goodness-of-fit tests, clustering, classification, or regression.

For publication P3 (see appendix A.3), I have used the Energy distance introduced by Rizzo and Székely [2016]. I have chosen this measure because it can be interpreted

intuitively for model comparison, it has good convergence properties and no density estimation is needed.

Energy Distance

Energy distance is a metric that measures the distance between two (possibly multivariate) probability distributions. It is called "energy distance" because of the analogy to the potential energy between objects [Rizzo and Székely, 2016]. It satisfies all axioms of a metric. The squared distance between the distributions $F(X)$ and $G(Y)$ is defined as

$$d^2(F,G) = 2\mathbb{E}||X-Y||_2 - \mathbb{E}||X-X'||_2 - \mathbb{E}||Y-Y'||_2 \geq 0, \tag{2.12}$$

with $\mathbb{E}$ being the expected value, $||\cdot||_2$ being the Euclidean norm, X and X' being independent and identically distributed (iid) variables, the same applies to Y and Y'.

The expected values can be implemented in a Monte Carlo framework as follows:

$$\mathbb{E}||X-Y||_2 = \frac{1}{N_{MC}^2} \sum_{k=1}^{N_{MC}} \sum_{l=1}^{N_{MC}} \sqrt{(\mathbf{x}_k - \mathbf{y}_l)^2}, \tag{2.13}$$

where $x_k \sim F$, $y_l \sim G$ and N_{MC} being the number of Monte-Carlo samples.

It can be interpreted intuitively as between-model variation ($\mathbb{E}||X-Y||_2$) and within-model variation ($\mathbb{E}||X-X'||_2$, $\mathbb{E}||Y-Y'||_2$).

Energy Score

In many real-world applications, the probability distribution cannot be estimated reasonably well based on limited measurement data. For such cases, probability metrics such as the Energy distance cannot be applied. However, so-called scoring rules address the issue of rating probabilistic models given deterministic measurements [Gneiting and

Raftery, 2007, Yao et al., 2018]. One such score is the so-called energy score, which is the counterpart of the energy distance [Ziel and Berk, 2019]. The energy score for the model predictive distribution G and observations $\mathbf{y}_{meas}$ writes as:

$$ES(G, \mathbf{y}_{meas}) = \frac{1}{2} E||Y - Y'||_2^\beta - E||Y - \mathbf{y}_{meas}||_2^\beta, \tag{2.14}$$

with $\beta \in (0, 2)$. In publication P3 (see appendix A.3), $\beta = 1$ was chosen as this is a standard choice for distributions that are not heavily tailed [Ziel and Berk, 2019].

Both quantities, the energy distance and the energy score, act on the same scale and can therefore be directly compared.

2.3.4 Visualizing Model Similarity

Based on energy distance and energy score, similarities between pairs of models and models and observations can be quantified. However, in the case of many models and many quantities of interest, an appropriate visualization technique is needed so that modelers can gain an intuitive understanding of the quantity of interest space.

Representing the similarities between N objects (models and observed data) leads to $n_{comb} = N \cdot (N-1)/2$ combinations. For each combination a distance is calculated, resulting in a symmetric $N \times N$ distance matrix. As the number of models to be compared can become high in extensive multi-model ensembles, visualization of the distances in two dimensions is not a straightforward task. Here, one has to handle the typical problem in dimensionality reduction methods: Balancing the interpretability with the preservation of the original structure [Liu et al., 2017]. I have tested various techniques and will present three of them that have proven to be suitable and have been used in publication P3 (see appendix A.3). Each method belongs to a different category of visualization methods: matrix-, axis- or hierarchy-based. Consequently, each visualization method has its particular strengths and weaknesses regarding a certain question we ask about the model set. Therefore, interpreting the different visualizations jointly gives a comprehensive understanding of the model set. The following overview is taken from [Schäfer Rodrigues Silva et al., 2022]:

1. Heatmaps: In a symmetric matrix, the distances between all pairs of objects are visualized through varying colors or intensity [e.g. Nandi and Sharma, 2020].

2. Radar charts: The axes are arranged radially starting from a common center. Each distance between two objects is plotted as a point on one axis. The resulting points are connected to a polygon, representing one object [e.g. Nandi and Sharma, 2020].

3. Dendrograms: These tree-like diagrams are typically used for visualizing hierarchical structures. A dendrogram consists of branches that connect objects depending on their similarity. The height at which two objects are joined together represents the distance between these objects [e.g. Nandi and Sharma, 2020].

While the first two methods are straightforward to implement, creating dendrograms requires several steps: In study P3 (see appendix A.3), the dendrograms have been created using an agglomerative hierarchical clustering approach [Xu and Wunsch, 2008] to assess the similarity of the model predictions.

In general, there are two types of clustering methods: partitional and hierarchical clustering. The first method groups data into a predefined number of clusters on a single level, so that each data point belongs to exactly one cluster. In contrast, hierarchical clustering groups data points into a nested, multi-level structure, so each data point belongs to a sub-cluster and associated higher-level clusters [e.g. Xu and Wunsch, 2008].

There are two ways of creating a hierarchical cluster: agglomerative (bottom-up) and divisive (top-down). The agglomerative method that I have used [Schäfer Rodrigues Silva et al., 2022] starts with N clusters, each containing one data point. Based on the distance matrix of these points, the algorithm identifies the pairs of clusters that have minimum distance and merges them. This merging procedure is repeated until all data points are finally in one overarching group [e.g. Xu and Wunsch, 2008]. The merging depends on the chosen linkage method, i.e. the definition of the distance between two clusters. Typical choices are single-linkage (or nearest-neighbor) that uses the smallest distance between data points in two clusters, complete-linkage (farthest-neighbor) that uses the largest distance between data points in two clusters, average-, median-, or centroid-linkage [Everitt et al., 2011].

3 Objectives and Contributions

The goal of this thesis is to demonstrate how a systematic assessment of model similarity can enhance multi-model methods. I consider the following steps to be part of such a systematic assessment:

1. Thorough selection of ensemble members considering structural similarities between the models.
2. Quantification of similarities between model predictions.
3. Intuitive visualization of similarities in a model set.

Available literature provides guidance about the first step, i.e the definition of a model set [e.g. Neuman, 2003, Refsgaard et al., 2012, Ferré, 2017, Enemark et al., 2019]. A commonly used method for defining a model set is the "varying complexity" approach [Enemark et al., 2019], i.e. starting from a detailed reference model, different model simplifications are defined (see Section 1.1). However, a systematic way to assess whether these simplified models are appropriate replacements for the reference model was lacking. From this gap in current research, I have derived the first research question of this thesis:

***RQ1**: How can we systematically assess how similar conceptually simplified model versions are compared to an original, more detailed model?*

This research question has been addressed in the publication P1 "Strategies for Simplifying Reactive Transport Models: A Bayesian Model Comparison" [Schäfer Rodrigues Silva et al., 2020] (see appendix A.1).

The method presented in P1 is based on the model justifiability analysis (see Section 2.1.3). This analysis involves many model evaluations and, thus, it becomes intractable for computationally expensive models. This leads directly to the next research question:

***RQ2**: How can we extend the similarity analysis so it is suitable for computationally expensive models?*

The necessary extension for computationally expensive models has been published in study P2 "Surrogate-based Bayesian Comparison of Computationally Expensive Models: Application to Microbially Induced Calcite Precipitation" [Scheurer et al., 2021] (see appendix A.2).

The research questions presented so far address step 1 of the suggested multi-model strategy. What remains open are steps 2 and 3 concerning the quantification and visualization of model similarities:

Model similarity has received considerable attention, especially in the field of climate modeling [e.g. Abramowitz, 2010, Knutti et al., 2010, Evans et al., 2013, Sanderson et al., 2015a,b, Abramowitz and Bishop, 2015, Knutti et al., 2017, Christiansen, 2018, Abramowitz et al., 2018]. Fewer studies suggested methods for its quantification [Abramowitz and Gupta, 2008, Sanderson et al., 2015a,b, Knutti et al., 2017] and visualization [Sanderson et al., 2015a,b]. However, none of these studies used a metric, i.e. a distance measure that fulfills certain properties (see Section 2.3.3) [Abramowitz and Gupta, 2008].

Against this background, the following research gaps exist:

***RQ3**: How can we visualize the similarities between probabilistic model predictions?*

This question has been addressed in publication P3 "Diagnosing Similarities in Probabilistic Multi-Model Ensembles - an Application to Soil-Plant-Growth-Modeling" [Schäfer Rodrigues Silva et al., 2022] (see appendix A.3).

The case studies vary across a wide range of areas in environmental system modeling to demonstrate the broad applicability of the developed methods and findings.

4 Results and Discussion

4.1 Strategies for Simplifying Models - a Bayesian Model Comparison

***RQ1**: How can we systematically assess how similar conceptually simplified model versions are compared to an original, more detailed model?*

A commonly used strategy modelers pursue to address model choice uncertainty is the "varying complexity" approach. This means all models in the set are based on the same concept, and the complexity is gradually increased or decreased. Though this strategy does not fulfill the MECE criterion (see Section 1.1), it is still helpful when it is not clear which level of detail is needed to model certain quantities of interest. Examples for studies that apply this strategy in environmental modeling are Schöniger et al. [2015], Hommel et al. [2016], Loschko et al. [2016], Brunetti et al. [2020].

Often, modelers wish to simplify a detailed model that has high computational costs and/or high parametric uncertainties due to a high number of processes it considers. To overcome these issues, modelers try to find simpler models that are still able to calculate the quantities of interest sufficiently well and hence can replace the original model. To define such simplified versions, processes or interactions are neglected or described by lower-order functions. This leads to the following questions:

> "Are the simplified models sufficiently unbiased and flexible enough to reproduce the computationally expensive reference model? Can we select a simplified model that represents the reference model best or discard a model that performs unsatisfyingly?"

[Schäfer Rodrigues Silva et al., 2020]

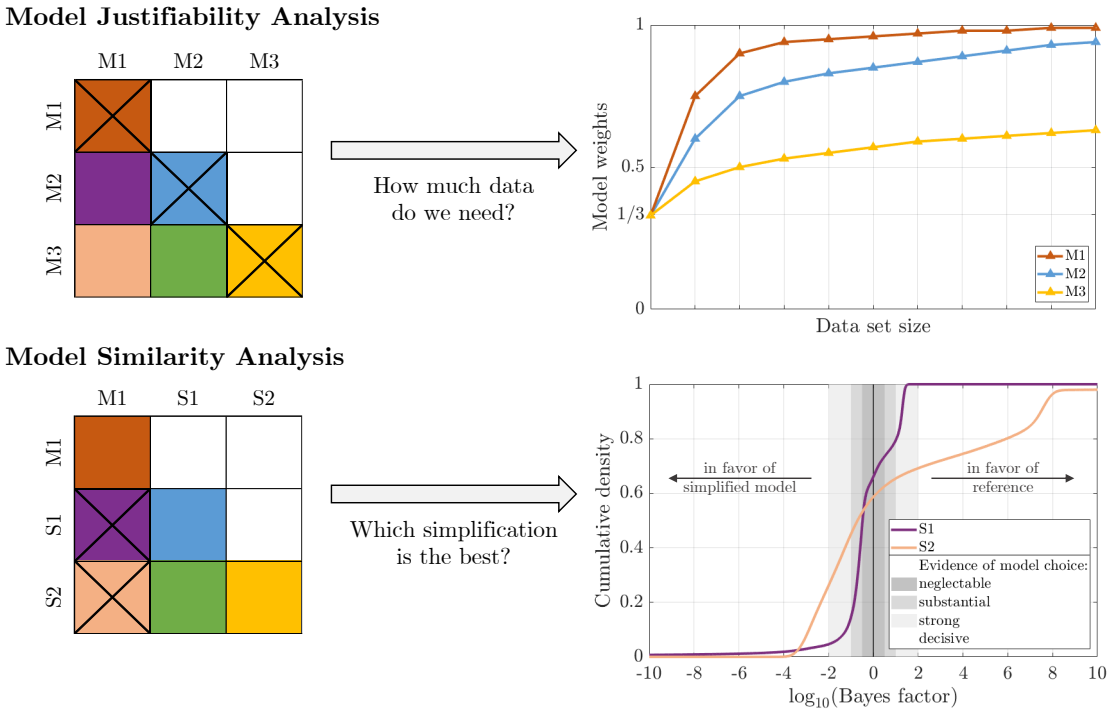


Figure 4.1: Schematic comparison of the model justifiability analysis and the model similarity analysis. Left: Model confusion matrices. The marked entries are the focus of the respective analysis. Right top: Average weights for the data-generating models over increasing data set size. Right bottom: cumulative distributions functions of the logarithmic Bayes factor for simplified models tested against a reference model (modified after Schäfer Rodrigues Silva et al. [2020]).

Quantifying Model Similarities Based on the Model Confusion Matrix

To address these questions, I have added a new way of interpreting the model confusion matrix introduced by Schöniger et al. [2015]. The original purpose of the method was to identify the justifiable level of model complexity given a specific amount and type of data. For my research question, however, this analysis has been interpreted from a new perspective, focusing on model similarity. To this end, the model confusion matrix has been used to quantify how similar the predictions of simplified models are compared to a reference model.

Figure 4.1 illustrates the different ways of analyzing a model confusion matrix: focusing on the main-diagonal entries to assess the model justifiability [Schöniger et al., 2015] or focusing on off-diagonal entries to assess model similarities [Schäfer Rodrigues Silva et al., 2020].

In the model justifiability analysis [Schöniger et al., 2015], each of the models in the set is sequentially treated as "synthetic truth" and the other models are tested against this "true" one. This yields a ranking of the models based on Bayesian model weights (see Section 2.1). These weights quantify the probability of the respective model to be the most appropriate one from the set. The main diagonal entries of the model confusion matrix are the self-identification weights, i.e. the weight each model receives when it generated the reference data itself. In the original analysis of Schöniger et al. [2015], these entries were used to assess whether a model's level of detail is justifiable given a certain amount of data. The off-diagonal entries 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. These entries can be used for finding the best simplification to replace the original model.

In the novel model similarity analysis, the most detailed model is considered as "true", and the simplifications are tested against it. In study P1 (see appendix A.1), these off-diagonal weights have been used to check how simplified models score through the eyes of BMS when the reference model generated the data. In addition to the analysis of the model weights, the decisiveness of the model choice has been assessed based on Bayes Factors (see Section 2.1).

Application to Reactive Transport Models

The method has been demonstrated using a set of five models that simulate aerobic respiration and denitrification in a heterogeneous aquifer at quasi-steady-state. The considered models are based on different conceptualizations and partly differ considerably in their computational costs.

1. The most complex reference model (M1) [Sanz-Prat et al., 2015] 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.
2. M2 neglects mixing due to dispersion. There are two sub-versions of this model:
   (a) uses the same parameters as M1.

   (b) uses adapted parameters to compensate for effects caused by neglecting dispersion.

3. M3 neglects biomass growth and decay [Sanz-Prat et al., 2015].

4. M4 replaces the advection-(dispersion)-reaction equation in Cartesian coordinates by the concept of cumulative relative reactivity solved along trajectories [Loschko et al., 2016].

5. M5 is also based on the cumulative relative reactivity approach and uses lower-order reaction kinetics: Aerobic respiration is described by zeroth-order decay and denitrification is modeled by first-order decay, instead of standard Monod kinetics as in M4.

The quantity of interest for assessing model similarities was flux-weighted nitrate concentrations in quasi-steady-state at different cross-sections. Clearly, the choice of the quantity of interest can influence the result of the similarity analysis. Therefore, the conclusions drawn from the analysis of the flux-weighted nitrate concentrations might change for other variables.

Key Findings and Implications

The analysis of the cumulative distributions functions (CDF) of the Bayes Factors in Figure 4.2 reveals that M2b, M4, and M5 are suitable candidates for replacing the reference model, while M3 is not a robust choice. M2a is virtually a duplicate of the original model in the chosen QoI space and, therefore, it might not be seen as a real simplification (for predicting this particular quantity of interest). Overall, the most simplified model M5 turned out to be the best replacement of the computationally expensive reference model as it scores well in the similarity analysis and tremendously reduces runtimes compared to the spatially explicit models (M1-M3) [Schäfer Rodrigues Silva et al., 2020].

The study has proven that the extended justifiability analysis [Schöniger et al., 2015] is a suitable tool for selecting the most appropriate simplification of a reference model. Apart from answering the first research question, the study also carries another interesting implication: When certain processes are neglected to find a simplified model description, it is often harder to define reasonable parameter prior distributions. This

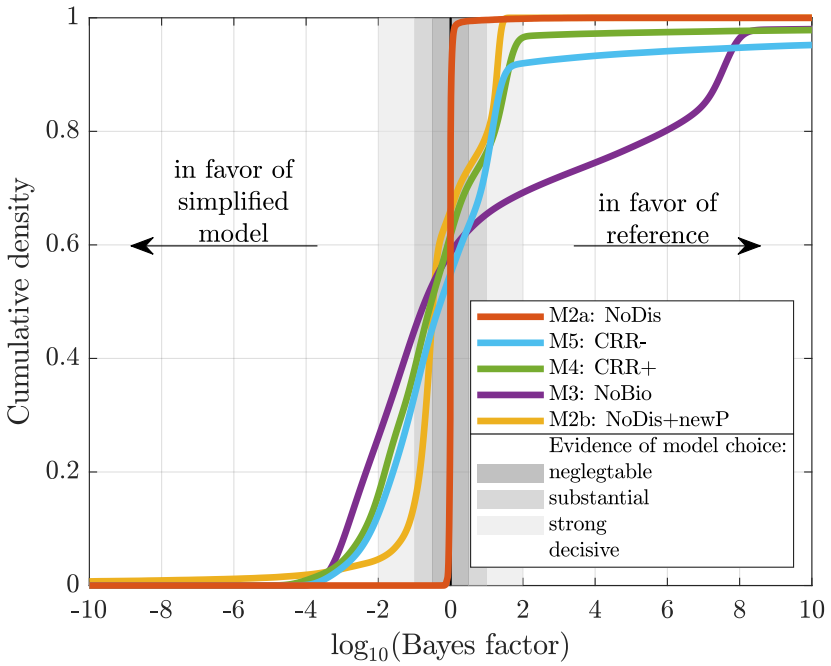


Figure 4.2: Cumulative distributions functions of the logarithmic Bayes Factor for M2–M5 tested against the reference model. Modified after Schäfer Rodrigues Silva et al. [2020].

is because, in the more detailed process description, the parameters can be interpreted physically and a plausible parameter distribution can be assigned accordingly. In the simplified process description, however, often effective parameters are used that compensate for the missing process description. Defining a reasonable distribution for effective parameters is often more difficult than for physical parameters [Schäfer Rodrigues Silva et al., 2020]. This finding has implications for the controversial definition of model complexity that has recently been discussed, e.g. by Höge et al. [2018] and Baartman et al. [2020]: Apart from considering the number of incorporated processes and interactions of a model or the prior/posterior shrinkage (Occam factor) [MacKay, 1992], it should also be considered how difficult it can be for modelers to a priori constrain effective parameters [Schäfer Rodrigues Silva et al., 2020].

4.2 Surrogate-Based Bayesian Comparison of Computationally Expensive Models

***RQ2**: How can we extend the similarity analysis so it is suitable for computationally expensive models?*

BME-based analyses such as BMS [Raftery, 1995] or the model justifiability analysis [Schöniger et al., 2015] typically involve many model runs in a Monte Carlo framework (see Section 2.1.3). For computationally expensive models, this soon becomes intractable. In such cases, surrogate models with negligible run times are commonly used to replace the original model (see Section 2.2).

Mohammadi et al. [2018] introduced a method that uses aPCE-based surrogate models for a BMS analysis, i.e. ranking models based on Bayesian model weights given observed data (see Equations 2.7 - 2.9). To account for the approximation error that is introduced when the surrogate is used instead of the original model, they introduced a correction factor. In the justifiability analysis, however, models are compared against each other instead of testing them against measurements. Consequently, the approximation of both models has to be considered. This has been achieved in my second publication "Surrogate-based Bayesian Comparison of Computationally Expensive Models: Application to Microbially Induced Calcite Precipitation" (Scheurer et al. [2021], see appendix A.2). The new method extends the method by Mohammadi et al. [2018] and enables the calculation of a model confusion matrix that represents the original models, but is calculated using their surrogates.

Two-stage Bayesian Model Analysis Based on aPCE Surrogates

In publication P2 (see appendix A.2), a two-stage Bayesian analysis has been carried out based on aPCE surrogates. In the first step, the classical BMS analysis has been used to rank the models based on experimental data [Mohammadi et al., 2018]. In a second step, the model justifiability analysis [Schöniger et al., 2015] has been used to gain insights into the data demand of the models. Combining both analyses allows a more sophisticated examination of the model set than the sole use of the classical BMS would give: It does not only identify the most appropriate model given the *currently*

available amount of measurements but also answers the question of whether this choice might change when more informative data became available in future [Schöniger et al., 2015].

After these two BME-based steps, which rank models based on a tradeoff between goodness-of-fit and simplicity, the model accuracy has been evaluated using the coefficient of determination R^2 between measured and predicted values. As this analysis only rates the model's accuracy independent of their complexity, it helps to interpret the BME-based results.

Application to Reactive Transport Models

The method has been demonstrated using a set of three models that simulate microbially induced calcite precipitation (MICP) and experimental data. MICP is used in engineering applications to alter the permeability of porous media [e.g. Hommel et al., 2015]. As described in Scheurer et al. [2021], 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.

Detailed descriptions of the models can be found in Hommel et al. [2015, 2016] and the experiment in Hommel et al. [2015] (see experiment "D1"). The conceptual differences between the models are listed in Table 4.1.

The analysis has been based on model predictions and measurements of calcium and calcite at different points in space and time. For each quantity of interest (calcite and calcium), a $d = 2$ order aPCE surrogate model has been built (as described in Section 2.2).

Table 4.1: Key differences of the investigated models. Adopted from Scheurer et al. [2021].

model	*full complexity*	*initial biofilm*	*simple chemistry*
simplifying assumption	-	pre-existing biofilm	precipitation determined by ureolysis
simulated time	3 203 460 s	3 109 860 s	3 203 460 s
biomass transport, 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

Key Findings and Implications

Figure 4.3 shows the model confusion matrices that have been calculated for the calcite content (top) and the calcium concentration (bottom) for increasing data set sizes. Recall that the main-diagonal entries of the model confusion matrices are the "self-identification weights", which represent the models' ability to identify their own predictions. A low self-identification weight indicates that the data set size is not sufficient for the model to identify its own predictions, or in other words, the model's level of detail is not legitimate given the respective data set. The self-identification weights clearly reflect the reduced data need of the simple chemistry model compared to its competitors.

The off-diagonal entries reflect the similarities between the models. They show that, even with the largest data set, the confusion between the *initial biofilm* model M_{IB} and the *full complexity* model M_{FC} remains strong, i.e. the models are similar regarding these quantities of interest.

The weights in the first column represent the models' appropriateness given the measurements. Except for the smallest calcite data set, these weights are smaller by orders of magnitude compared to the values between the models. This indicates that there is at least one relevant process missing in all models as presumed by Hommel et al. [2015]. Considering that the "true" model is not even assumed to be part of the model set and the missing process is unknown, the current problem can be assigned to the so-called "$\mathcal{M}$-open" view of model settings [Vehtari and Ojanen, 2012]. In contrast to this, the "$\mathcal{M}$-closed" view [Bernardo and Smith, 2009] represents a scenario in which the data-generating process is assumed to be in the model set, and "$\mathcal{M}$-completed" means that

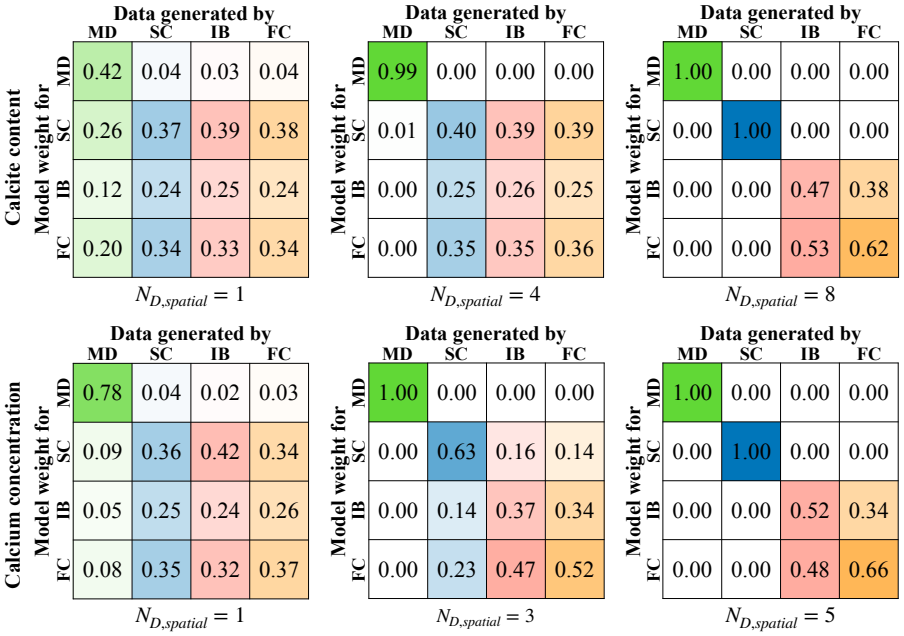


Figure 4.3: Model confusion matrices for the calcite content (top) and the calcium concentration (bottom) for increasing data set sizes. Adopted from Scheurer et al. [2021].

the data-generating process is not in the model set but can be approximated.

Technically speaking, applying BMS to an "M-open" problem violates the assumption of the method that one of the models in the set is the true one [Höge et al., 2019]. However, in real-world applications, this assumption is often relaxed to the task of identifying the *most appropriate* model in the set [Höge et al., 2019]. From this perspective, the results of the BMS analysis (Figure 4.4) reveal which model is the most appropriate given the current set, but a weight of 100% does not mean that the true data-generating process is identified.

In the last step, the coefficient of determination R^2 has been used to evaluate whether a model ranks well in the BMS analysis because of its simplicity or its goodness-of-fit, or a combination of both. For both quantities of interest, the *simple chemistry* model M_{SC} has the highest goodness-of-fit. As it has shown the smallest data demand in the justifiability analysis, it can be concluded that its good BMS ranking results from both simplicity and goodness-of-fit. Contrarily, the *initial biofilm* model M_{IB}, scores worst in the BMS analysis because of its high level of complexity (see model confusion matrices in Figure 4.3) and a worse fit to the measurement data. Even though this model is more similar to the full complexity model than the simplified chemistry model,

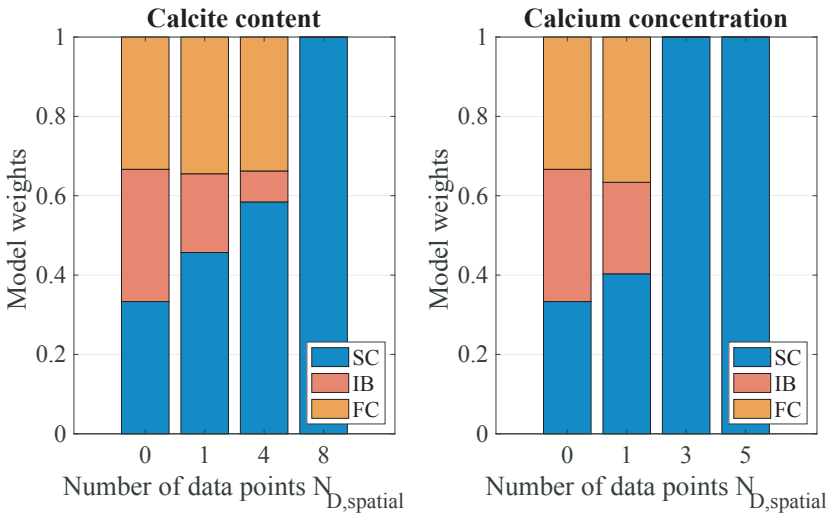


Figure 4.4: Average model weights over increasing data set size. Left: calcite content, right: calcium concentration. Adopted from Scheurer et al. [2021].

the latter turns out to be the better replacement.

Overall, the analysis has shown that the simplified chemistry model is the best choice for the given quantities of interest, however, it is only the best among an imperfect model set. On the methodological level, the new surrogate-based two-stage Bayesian analysis (BMS and justifiability analysis) has proven to be a suitable tool for assessing computationally expensive model sets regarding their data demand and goodness-of-fit.

4.3 Quantifying and Visualizing Similarities in Probabilistic Multi-Model Ensembles

***RQ3**: How can we visualize the similarities between probabilistic model predictions?*

Though it had been widely recognized that the similarities between the individual models play a crucial role for an effective use of multi-model ensembles, no method existed that intuitively visualizes these similarities and thus guides the choice of the ensemble members.

Table 4.2: Overview of the methods used for comparing models and observations in publication P3.

Comparison of	probabilistic predictions	probabilistic predictions - noisy measurements	probabilistic predictions - deterministic measurements
Method	Energy distance	Energy distance	Energy score

Quantifying Model Similarities Based on Energy Statistics

I have addressed this gap in publication P3 (see appendix A.3) by showing how an existing statistical method (so-called energy statistics) can be used to quantify the similarity between models and how these similarities can be visualized in a way that gives modelers an intuitive understanding of the model set. The models have been compared based on their predictions. This approach has the advantage that the goodness-of-fit to observations can be assessed with the same method and modelers can compare the similarities between two models and between models and measurements on the same scale (see Section 2.3.3). Table 4.2 gives an overview of the methods used for comparing models and observations.

As previous studies pointed out, it can be misleading to compare calibrated models because often non-physical parameters are used, which compensate for structural deficiencies of the model [e.g. Vogel and Sankarasubramanian, 2003, Wallach et al., 2020, Wallach, 2011]. As a consequence, a model might have a good fit to measurements of a certain variable, but a poor one for variables it has not been calibrated on. Therefore, I have followed the recommendation of Vogel and Sankarasubramanian [2003] and have analyzed the models based on their prior predictive distributions.

Based on these distributions, I have used the energy distance and the energy score to construct distance matrices. To ease the interpretation of the results, three different visualization techniques (heatmaps, radar charts, and dendrograms, see Section 2.3.4) have been used. Each method highlights a particular aspect of the model set and, therefore, a combined interpretation facilitates the further model development process [Schäfer Rodrigues Silva et al., 2022].

Application to Soil-Vegetation-Atmosphere Models

The method has been demonstrated using a set of three models that simulate the soil-vegetation-atmosphere continuum. The models are CERES [Ritchie et al., 1988], SUCROS [van Laar et al., 1997], and SPASS [Wang and Engel, 2000, Gayler et al., 2002], which are part of the multi-model library Expert-N [Priesack, 2006]. The quantities of interest are two in-season variables, phenology (BBCH) and leaf area index (LAI), and the end-of-season variable grain yield.

Key Findings and Implications Regarding the Visualization Methods

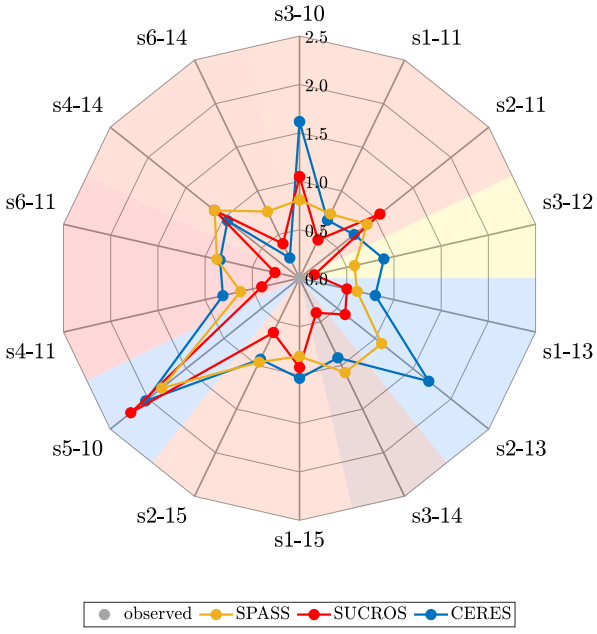


Figure 4.5: Radar chart showing the energy distance between all models and the observations based on yield. Each colored line represents the distance of one model to the observations. Each edge of the net represents one site-year, i.e. a different instance of the QoI space. Segment colors resemble the annual weather conditions: hot and dry (red), average (yellow) to cold and wet (blue).

Before discussing what the results imply for the Expert-N model set, I first want to focus solely on the different visualization methods: In Figure 4.5, a radar chart is equipped with a color-coded background that shows warm and dry years in red, cold and wet years in blue and average years in yellow. This representation facilitates the

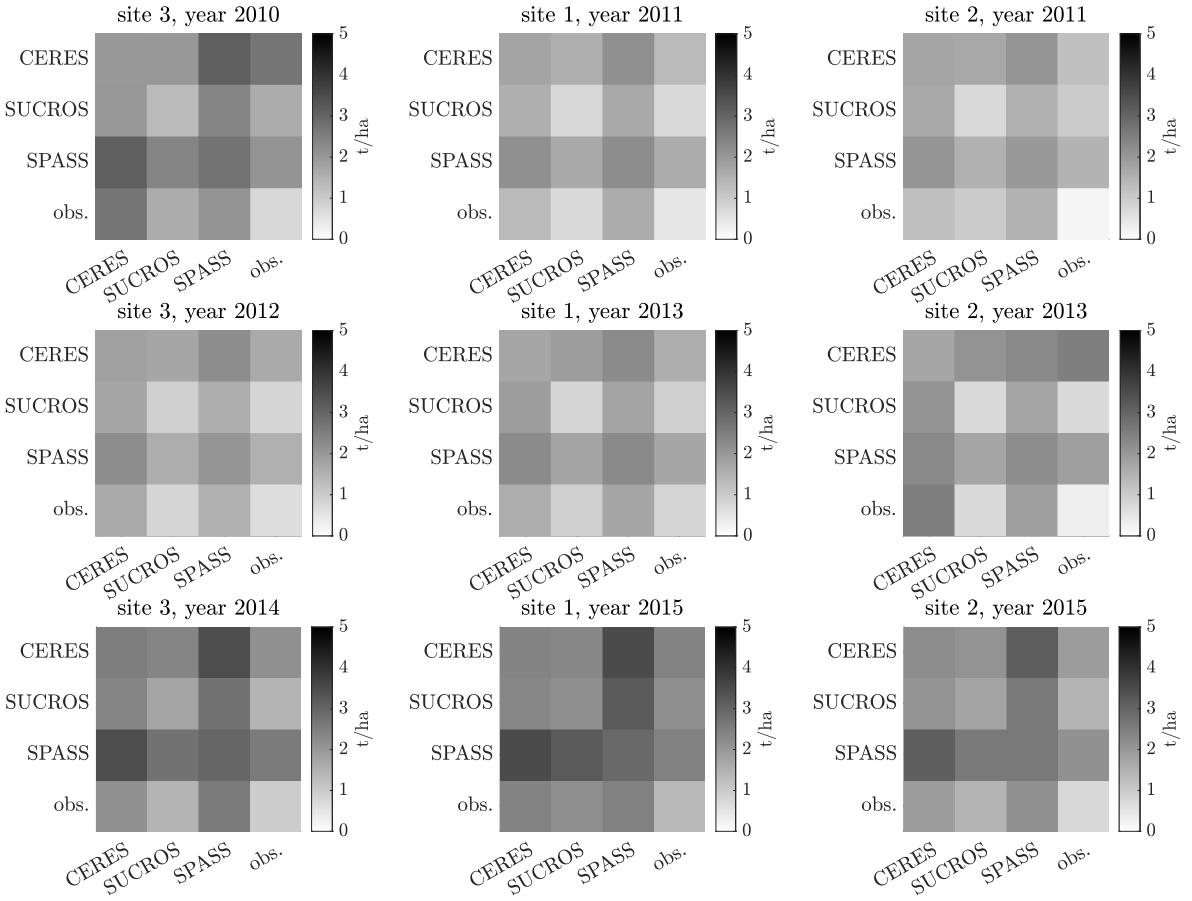


Figure 4.6: Heatmaps reflecting the similarities between models and observations based on yield predictions for site 1-3. The color-coding represents the values of the individual components of the energy distance: $\mathbb{E}||X - Y||_2$ (off-diagonal entries) and $\mathbb{E}||X - X'||_2$ (main diagonal entries).

interpretation of the models' behavior under specific conditions. The chart is centered on the observations and, hence, the distances from the center can be interpreted as the goodness of fit to measurements. With this visualization, the data can be represented densely for all site-years. However, to show all the distances between N objects (models and data), $N-1$ radar charts are needed. Therefore, this method is well suited to check whether the similarities between one specific object (e.g. measurement data or a certain reference model) change under varying conditions.

In contrast to the other visualization methods, the heat maps in Figure 4.6 are not based on the energy distance, but on its components $\mathbb{E}||X - X'||_2$ (main diagonal entries) and $\mathbb{E}||X - Y||_2$ (off-diagonal entries). Thereby, the heat maps do not only show the similarities between pairs of models or models and measurements, but also

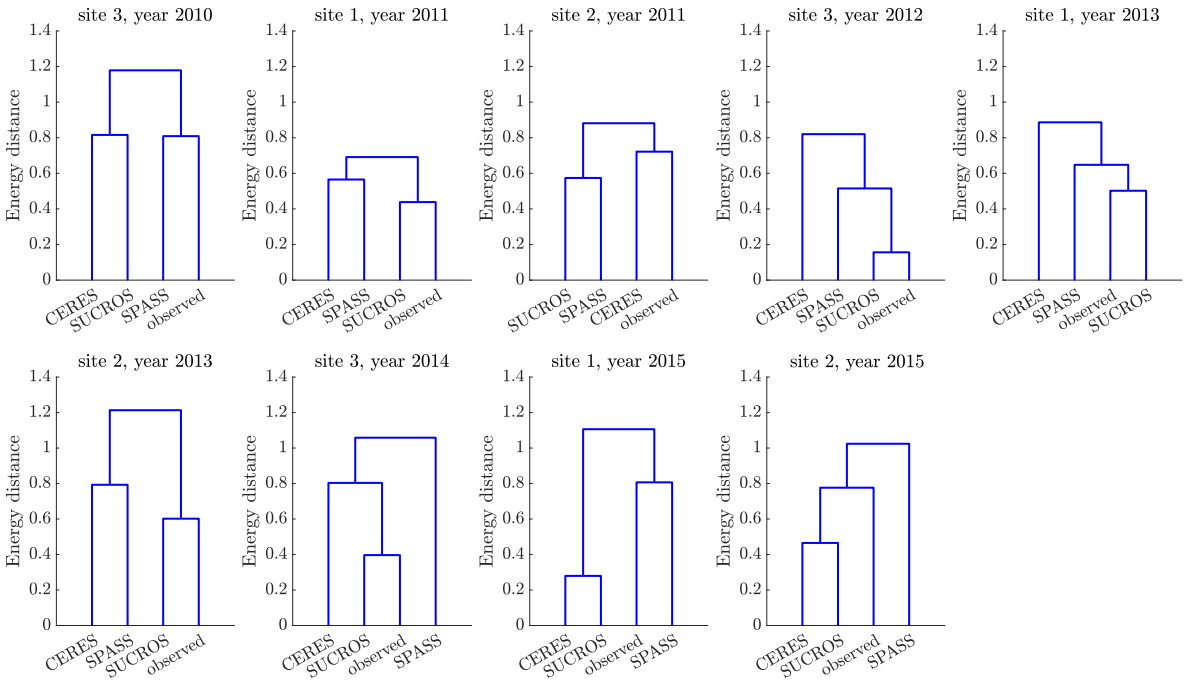


Figure 4.7: Dendrograms based on the energy distance between models and observations (yield predictions) for sites 1-3.

the within-model or within-measurement variation. Heat maps allow the visualization of many objects at once. Here, for each site-year one heat map is used. However, one could also think about merging the maps for all site-years into a single one, representing each model for each site-year as one row and column. Also, the color-coding of weather conditions as done in the radar charts could be done here: One can use different colors in the heat maps for coding different conditions and varying intensity for coding similarities. However, test results (not shown here) have shown that both extensions make the plot overloaded and the information becomes hard to capture.

Figure 4.7 shows dendrograms that visualize the model set and the measurements in hierarchical structures. Similar objects are joined together in an overarching group. The height at which two objects are joined represents the distance between them. Dendrograms are well suited for comparing a large number of objects and identifying clusters. In the context of multi-model ensembles, such clusters can identify settings as "all models cluster together far-off from measurements" and help modelers to judge into which category defined by Bernardo et al. [1999] the modeling problem falls: $\mathcal{M}$-open, $\mathcal{M}$-completed, or $\mathcal{M}$-closed (see Section 2.3.1).

Table 4.3: Comparison of the visualization methods. The checkmark means that the method is well suited, the checkmark in parentheses means that the method can be used for the task in certain cases, but the visualization might become overloaded. Adopted from Schäfer Rodrigues Silva et al. [2022]

Visualization Method	Radar Charts	Heat Maps	Dendrograms
Type	axis based	matrix based	hierarchy based
Comparison of many objects	✗	✓	✓
Comparison of many conditions	✓	(✓)	✗
Easy identification of Clusters	✗	(✓)	✓
Color-coding possible	✓	(✓)	(✓)
Variation within and between objects	✗	✓	✗

Table 4.3 summarizes the features of the visualization methods and how well they are suited for certain use cases. As highlighted by Liu et al. [2017], visualizing high-dimensional data sets usually requires the joint use of multiple techniques to explore all relevant aspects.

Key Findings and Implications Regarding the Expert-N Model Set

The radar chart in Figure 4.5 shows that SUCROS is closest to the measurements for most site-years. The color-coded background (red: warm and dry, yellow: average, blue: cold and wet) of the radar chart does not reveal any dependency of the models' performances on the weather conditions. The same holds for model similarities (figure not shown here, please refer to Figure 4 in P3, appendix A.3).

Figure 4.8 shows time series of the within-model spread (top), the energy distance between pairs of models (middle), and the energy score between models and observations (bottom). Subfigure (a) is based on phenology (BBCH) data and subfigure (b) is based on leaf area index (LAI) data. In the analysis of all site-years, SPASS shows the best goodness-of-fit regarding in-season variables LAI and BBCH (see P3, appendix A.3). However, its yield predictions have very high variance and, hence, it scores worse in this ranking.

This result confirms previous studies, which emphasized that it is often not the same model that performs best in predicting different quantities of interest [Hagedorn et al., 2005, Palosuo et al., 2011, Martre et al., 2015].

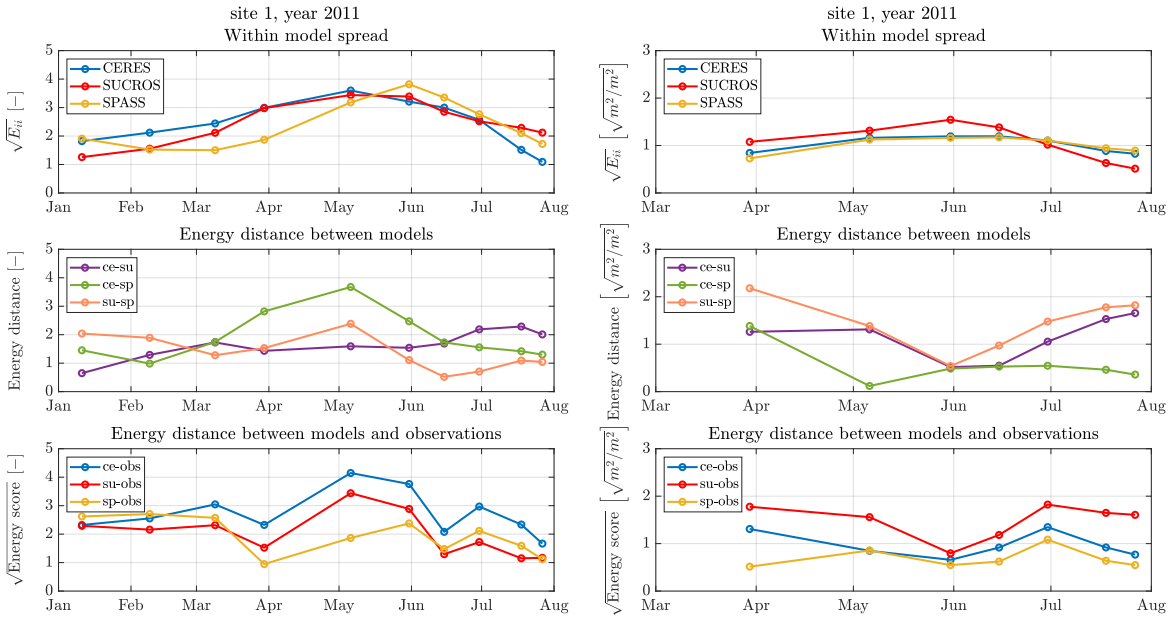


Figure 4.8: Top: Within model spread (square root of the mean Euclidean distance between the samples within each model $\sqrt{E_{ii}}$). Middle: Energy distance between pairs of models. Bottom: Energy score between models and observations. (a) phenology, (b) leaf area index.

Also, the similarities between models (see middle in Figure 4.8) depend on which variable is analyzed: While the LAI predictions of CERES and SPASS are the most similar ones and the predictions of SUCROS and SPASS are the most dissimilar ones, the opposite is true considering the yield predictions.

The analysis has revealed that each model has its strengths and weaknesses and similarities between models also differ depending on the considered variable. Therefore, a combined ensemble prediction might give more robust results.

5 Conclusions and Outlook

As stated in the introduction, model choice uncertainty is a major source of uncertainty in modeling environmental systems. In many fields of science and engineering, multi-model methods are used to address this type of uncertainty. These methods can be beneficial in many ways:

1. They can raise awareness of how much simulation results differ depending on the conceptual model.
2. They can yield more robust predictions.
3. They can help to find a level of model complexity that is appropriate given the amount of available calibration data.

In this project, my overall goal was to enhance multi-model methods by accounting for model similarities. I have broken down this goal into three research questions. In the following, I will take up and answer these questions.

RQ1: How can we systematically assess how similar conceptually simplified model versions are compared to an original, more detailed model?

In many applications, it is not clear which level of detail is needed to model a certain process. A common strategy to address this issue is to define simplifications of a reference model. Depending on the quantity of interest and the modeling objective, the simplifications might have more or less influence on the results. So how can we decide whether a simplified model replaces the more detailed reference appropriately? In study P1 [Schäfer Rodrigues Silva et al., 2020], I have shown that the model confusion matrix [Schöniger et al., 2015] can be used beyond its original purpose to answer this

question. To this end, Bayesian model weights are calculated based on the prior predictive distributions of the competing models. Considering the predictions of the most detailed model as a reference, the Bayesian model weights quantify the probability that the simplified models have generated these reference data. In this way, the Bayesian model weights can be seen as a probabilistic measure of model similarity and hence as a tool for finding the best simplification of a reference model.

RQ2: How can we extend the similarity analysis so it is suitable for computationally expensive models?

The method that I used to answer RQ1 involves many model evaluations in a Monte Carlo framework (see Equation 2.3). Therefore, it is infeasible for computationally expensive models. This drawback directly led to RQ2, which was answered in P2 [Scheurer et al., 2021]: Surrogate models approximate the original, computational expensive models and have negligible runtimes. Based on such surrogate models, the Bayesian model weights can be calculated. However, as we want to quantify the similarity between the original models, the errors introduced by the approximation of these models with the surrogates must be considered. This has been achieved by a newly introduced correction factor, which enables us to quantify similarities of the original models based on their surrogates. By this extension, it is now possible to calculate the model confusion matrix for computationally expensive models.

RQ3: How can we visualize the similarities between probabilistic model predictions?

Though it has been widely recognized that model similarities are important for multi-model methods, no technique existed that could provide modelers an intuitive visualization of their model set. I have addressed this issue and used the so-called energy distance for comparing probabilistic predictions of two models. With the same method, the models' goodness-of-fit to noisy measurements can be quantified and, with the related energy score, it is possible to quantify the similarity between probabilistic and deterministic data sets. Hence, energy statistics are widely applicable. Moreover, this method is easy to implement and has good convergence properties, so fewer samples

are needed compared to the Bayesian analysis that was applied in P1 and P2. For visualizing the resulting model distances, I have used different visualization techniques that were axis-, matrix- or hierarchy-based. It has been demonstrated that which method is best suited depends on the size of the model set and whether this set is to be analyzed under many different conditions. As each visualization method highlights a particular aspect of the model set, a combination of different methods should be used to gain a comprehensive picture.

The methods that I have developed in the course of this thesis can enhance multi-model methods, as they...

1. allow modelers to assess whether simplified models are a good replacement for a more detailed reference;
2. are widely applicable, even for computationally expensive models;
3. give modelers an intuitive understanding of the model set, and hence, help to find a multi-model strategy that best suits the QoI space setting at hand.

Ideas for Future Work

During this project, I came across many interesting studies from a wide variety of research fields. Even though it is tempting to delve into all these topics, I had to narrow down the scope of this project. Of course, many of these promising approaches remained in the back of my mind, and so I would like to mention them here and thus provide possible starting points for future work.

Uncertainty visualization has been considered a "top scientific visualization research problem" [Johnson, 2004]. Conceptual uncertainty is one of the key sources of uncertainty in geoscience and yet it is often underrepresented in the visualization and thus in the communication of uncertainties. Therefore, future work should further investigate visualization tools that might be helpful to address this issue in environmental modeling. Examples of promising tools are: (1) data context maps [Cheng and Mueller, 2016]: In a single "map", distances between samples and between variables are shown. This allows modelers to explore the relationship between samples and variables simultaneously, which was not possible in classical ways of visualizing data matrices. However, when interpreting these maps, it is important to keep in mind the projection-induced

distortion. (2) Uncertainty aware PCA [Görtler et al., 2019] offers a way to reduce the dimensionality of a data set accounting for uncertainties of the input variables.

Besides the question of how to visualize conceptual uncertainty, also the analysis of multivariate model predictions and measurements could be approached from a different perspective: A non-parametric way of comparing high-dimensional data sets is the so-called "depth statistics". This is based on the proposal of Tukey [1975] to generalize the median to the "deepest point" in a high-dimensional point cloud [Mosler, 2013, Singh and Bárdossy, 2012]. Depth-based statistics could be used for assumption-free quantification of similarities between multivariate data sets. Another emerging field that infers information about point clouds based on their shape is topological data analysis (TDA) [Carlsson, 2009, Chazal and Michel, 2017].

In this work, I have quantified similarities between models based on their predictions. The advantage thereof is that we can embed both, models and observations, in a common "quantity-of-interest (QoI) space" and use suitable visualization techniques to get a better understanding of the model set at hand. The proposed visualizations could be useful for judging in which of the settings defined by Bernardo et al. [1999] the modeling problem can be allocated in: $\mathcal{M}$-open, $\mathcal{M}$-completed, or $\mathcal{M}$-closed (see Section 2.3.1). Future research might explore how visualizing the QoI space can then guide the choice of an appropriate multi-model method and the development of alternative models.

To be able to judge whether the model set is an actual "team of rivals" [Ferré, 2017], assessing the structural similarity seems to be necessary. The information theory-based method presented by Bennett et al. [2019] could be a suitable tool. However, test results that have not been shown here indicate that very large data sets are needed for a successful application of this technique. Therefore, future work could tackle this issue and compare models based on both, their structure and their predictions. I expect a joint interpretation of both types of similarities to be highly valuable for further improving multi-model approaches.

A Publications

A.1 Strategies for Simplifying Reactive Transport Models - a Bayesian Model Comparison

AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR028100

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

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

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).

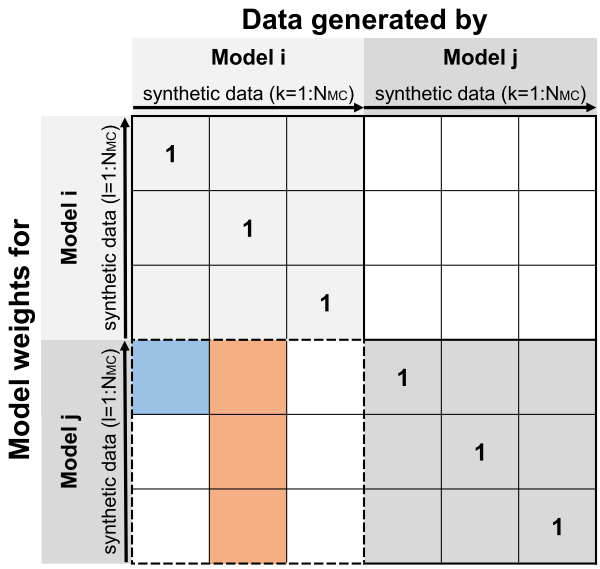


Figure 1. Schematic illustration of the model confusion matrix for two models, *i* and *j*. Blue box: likelihood of a single realization drawn from Model *j* given a realization drawn from Model *i*. Red box: BME value (average likelihood) of Model *j* given a single realization *k* of Model *i*. This 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 $P\left(M_i\right)$ 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 $P\left(M_i\right) = \frac{1}{N_m}$ that treat all models in the set as equally likely.

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

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

in which $p\left(\mathbf{y}_0|M_i\right)$ 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):

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

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 $p\left(\mathbf{u}_i|M_i\right)$ 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):

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

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 $p\left(\mathbf{u}_i|M_i\right)$. 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 $p\left(\mathbf{y}_{0,ik}|M_j, \mathbf{u}_{jl}\right)$ 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)
*M*1	spatially explicit	dispersion, dynamic biomass	10	38.7
*M*2	spatially explicit	dynamic biomass (Monod)	10	33.3
*M*3	spatially explicit	dispersion	10	31.3
*M*4	streamline based	cum. rel. reactivity (Monod)	5	0.1
*M*5	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 ($\text{mol}_C/\text{mol}_i$) 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 $\mathbf{x}$ 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 $F\left(\mathrm{T}\right)$ 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:

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

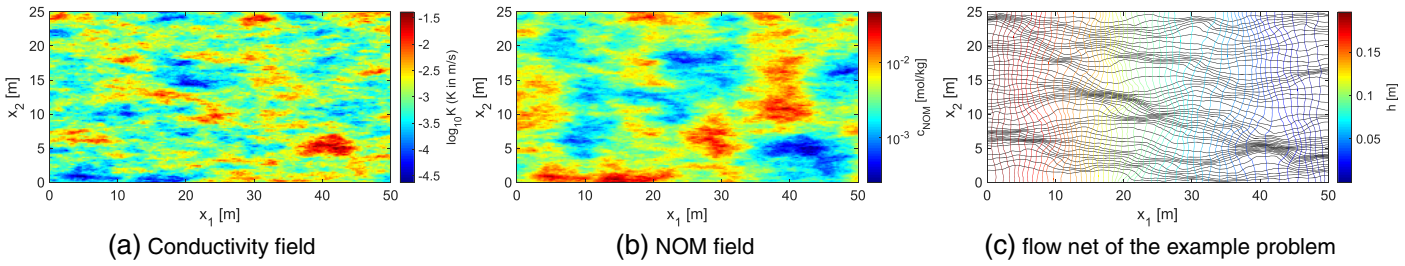


Figure 2. (a) Conductivity field, (b) NOM field, and (c) flow net of the example problem.

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

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 $\sigma^2(c) = 0.01\,\text{mol}^2/\text{L}^2$ 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 $\sigma^2(\bar{c}) = \frac{\sigma^2(c)}{n_{st}} + \frac{n_{st}-1}{n_{st}}\rho\sigma^2(c)$. This results in a slightly decreased variance related to the measurement error of the mean concentration $\sigma^2(\bar{c}) = 0.0092\,\text{mol}^2/\text{L}^2$.

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.

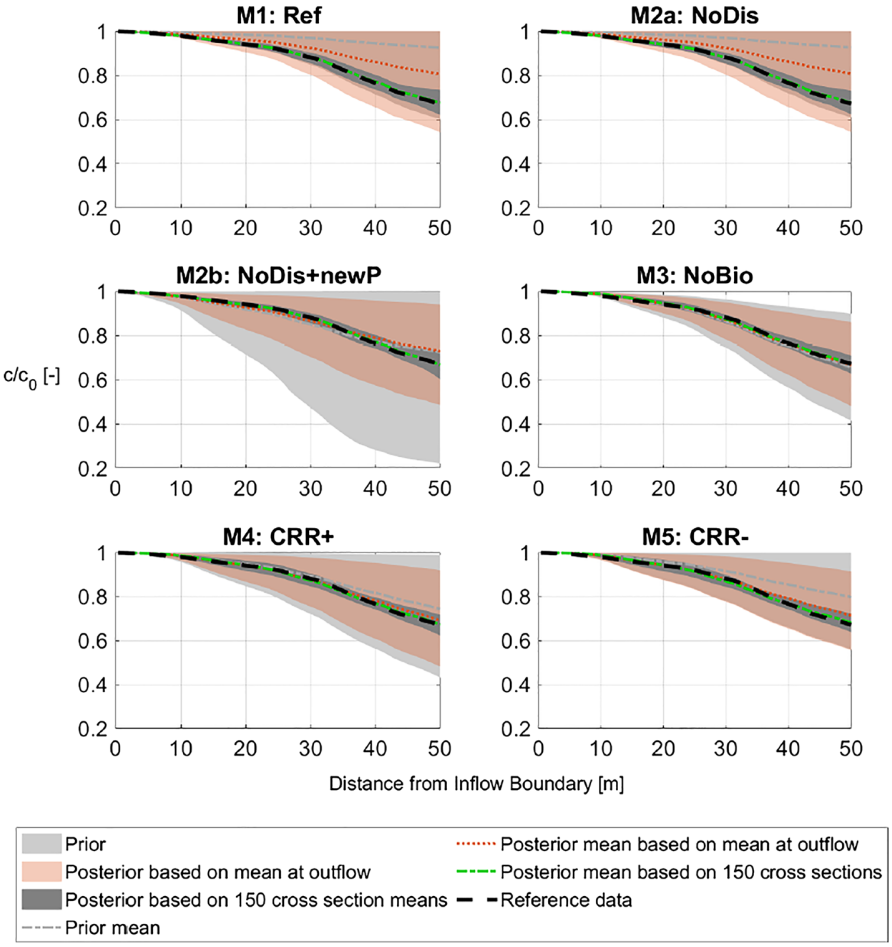


Figure 3. Ninety percent credible intervals for prior state (light gray), posterior after calibrating on one data point at the outflow (red), and posterior after calibrating on 150 equidistant data points (dark gray). Prior means (light gray line), posterior means after calibrating on one data point 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.

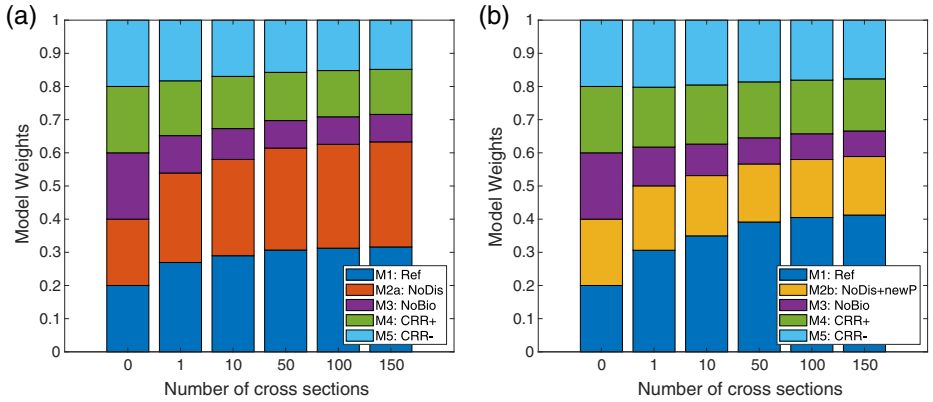


Figure 4. Average model weights over increasing data set size, when M1 generates the data. (a) Using the same parameter distributions for M1 and M2. (b) Testing parameter distributions that compensate for the missing dispersion in M2.

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.

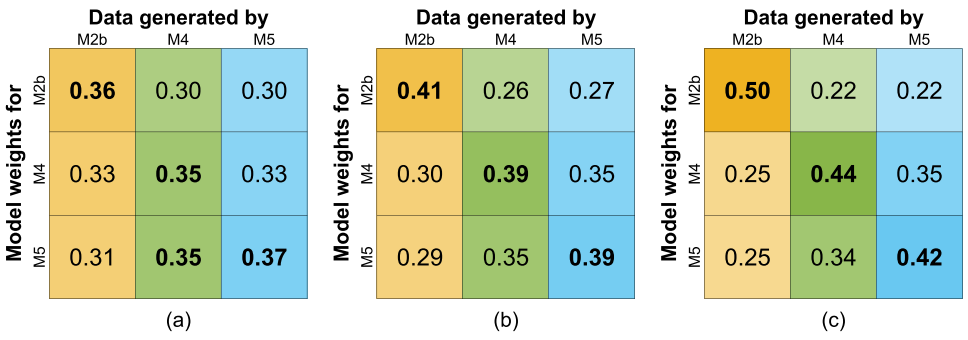


Figure 5. Model confusion matrices for M2b, M4, M5 based on concentrations in (a) a single cross section and (b) 25 and (c) 150 cross sections. Columns refer to the models that generate the data, rows to the models that we calculate the weights for. The highest weight for each data generating model (column)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

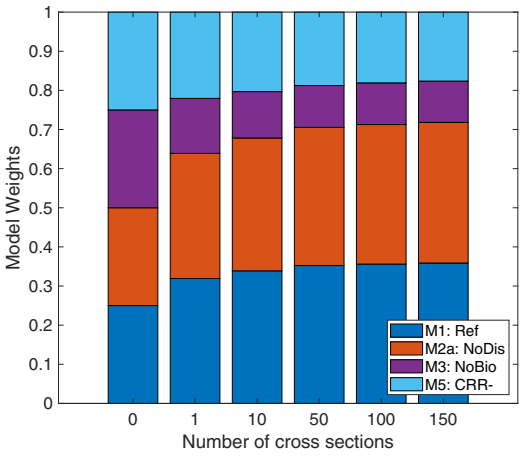


Figure 6. Average model weights when M1 generates the data over increasing data set size for the reduced model set.

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).

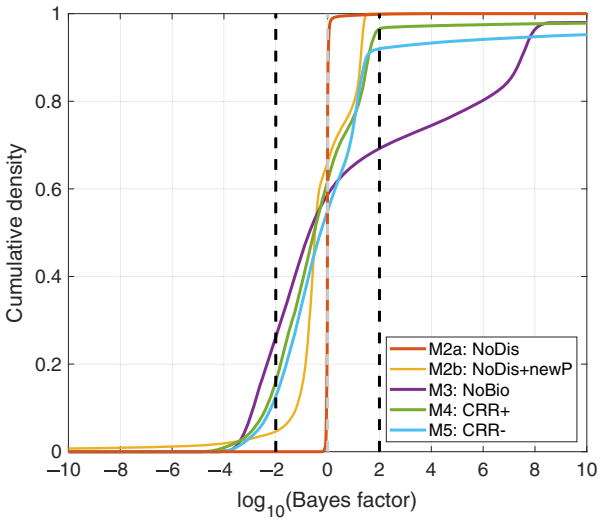


Figure 7. CDFs of the logarithmic Bayes Factor for M2–M5 tested against the reference Model M1.

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.
$c^{O_2}_{mob}$	Dissolved oxygen (mobile phase)	0 mol/L	$2.5 \cdot 10^{-4}$ mol/L
$c^{NO_3^-}_{mob}$	Nitrate (mobile phase)	0 mol/L	$1 \cdot 10^{-4}$ mol/L
$c^{CH_2O}_{mob}$	Dissolved organic carbon (mobile phase)	$3 \cdot 10^{-4}$ mol/L	0 mol/L
c^{bac}_{immob}	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 ($\text{mol}_{O_2}^{bac}/\text{mol}_C$)
$Y_{NO_3^-}$	Yield coeff. of nitrate	0.25 ($\text{mol}_{NO_3^-}^{bac}/\text{mol}_C$)
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 (μmol_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	$unif.(a,b)$	$a = 0.2, b = 0.3$	($\text{mol}_{O_2}^{bac}/\text{mol}_C$)
$Y_{NO_3^-}$	Yield coefficient of nitrate	$unif.(a,b)$	$a = 0.2, b = 0.3$	($\text{mol}_{NO_3^-}^{bac}/\text{mol}_C$)
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

A.2 Surrogate-based Bayesian Comparison of Computationally Expensive Models: Application to MICP

Computational Geosciences
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

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.

Published online: 05 August 2021

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:

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

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:

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

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 $\tilde{M}_{\setminus \boldsymbol{\omega}^{(i)}}\left(\boldsymbol{\omega}^{(i)}\right)$ 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 $P(M_k)$ are updated to posterior model weights $P(M_k|\boldsymbol{y}_0)$ using Bayes' theorem:

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

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 $P(M_k) = \frac{1}{N_\text{m}}$ 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

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

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:

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

where $\boldsymbol{R}$ is the covariance matrix of the measurement error ϵ of size $N_\text{s} \times N_\text{s}$ (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:

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

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:

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

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

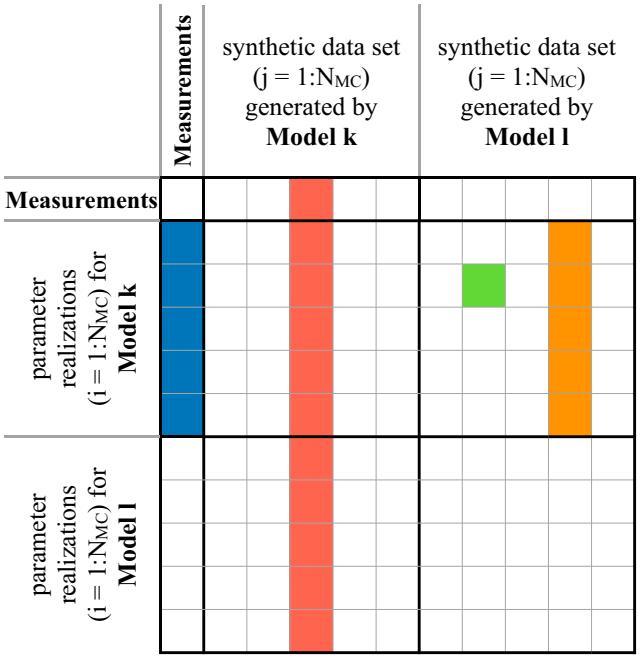


Fig. 1 Schematic illustration of constructing the model confusion matrix

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

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

or

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

with

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

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:

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

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:

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

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]:

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

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

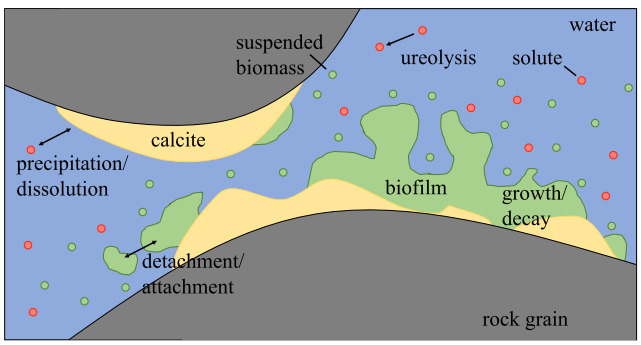


Fig. 2 Schematic view of relevant processes and phases during MICP after [25]

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.

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).

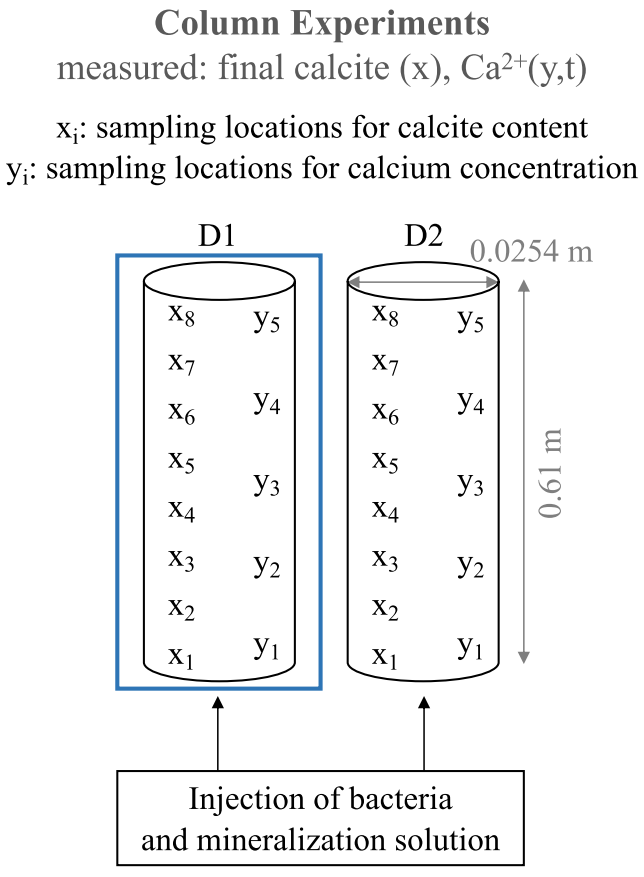


Fig. 3 Column experiment setup by [25] with measurement locations for calcite content and calcium concentration with analyzed column D1

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

after pulse \ pulse number	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 $c_{\mathrm{a},1}$, $\left[s^{-1}\right]$
- the coefficient for attachment to arbitrary surfaces $c_{\mathrm{a},2}$, $\left[s^{-1}\right]$
- 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
$c_{a,1}$	$[1 \cdot 10^{10} s^{-1}, 1 \cdot 10^{-7} s^{-1}]$
$c_{a,2}$	$[1 \cdot 10^{10} s^{-1}, 1 \cdot 10^{-6} s^{-1}]$
ρ_f	$[1\ \mathrm{kg/m^3}, 15\ \mathrm{kg/m^3}]$
k_{ub}	$[1 \cdot 10^{-5}\ \mathrm{kg/kg}, 5 \cdot 10^{-4}\ \mathrm{kg/kg}]$

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

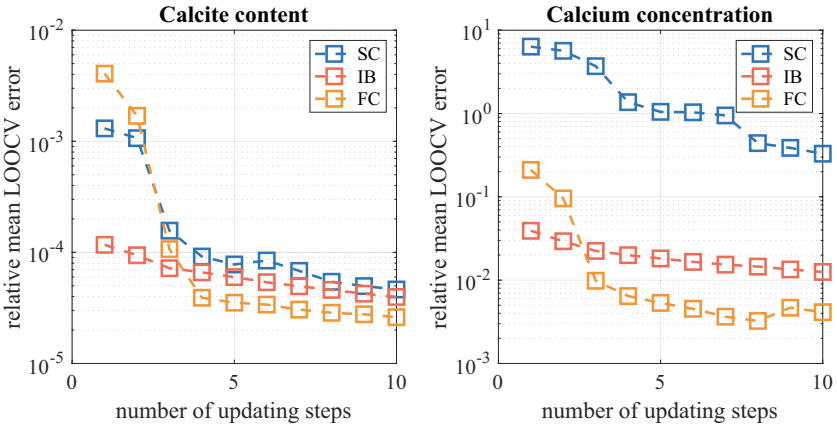


Fig. 4 Relative mean LOOCV errors for calcite content and calcium concentration with increasing number of updates

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

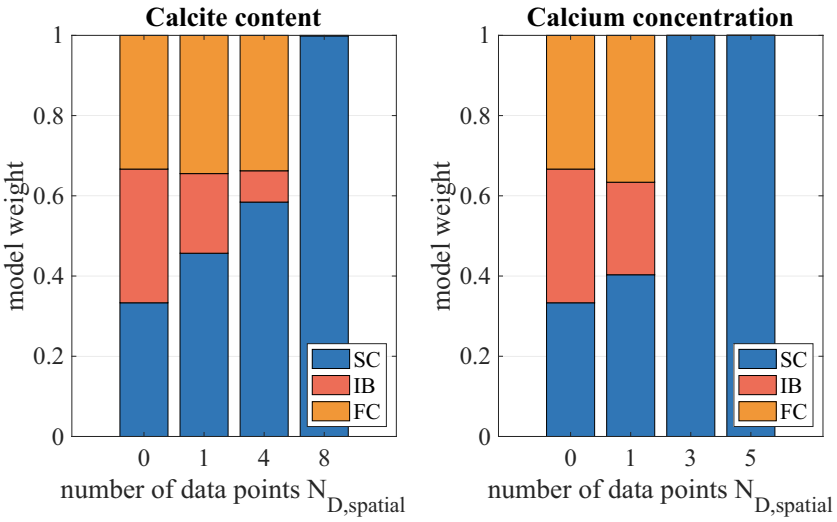


Fig. 5 Model weights for the prediction of calcite content and calcium concentration over increasing amount of used spatial data points $N_{D,\text{spatial}}$

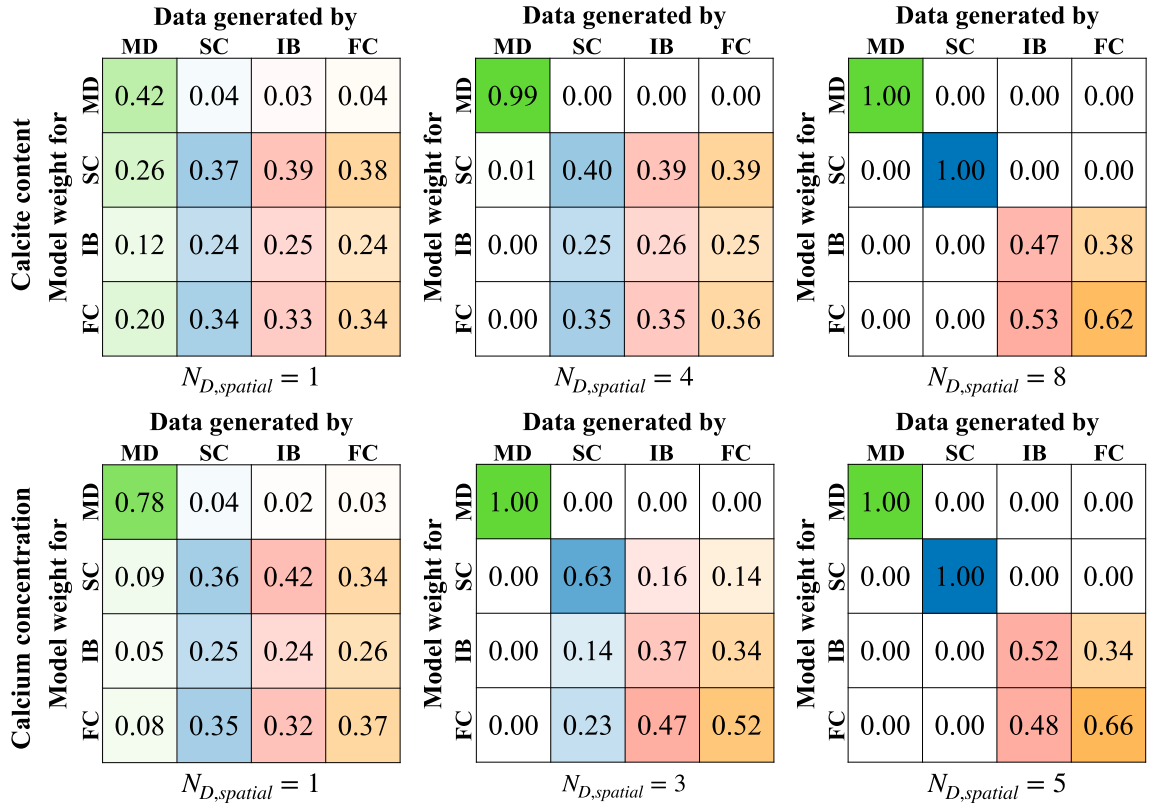


Fig. 6 Model confusion matrices for calcite content [%] and calcium concentration [mol/m^3] of the three models and the measurement data (MD) over increasing amount of used spatial data points $N_{D,\text{spatial}}$

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,

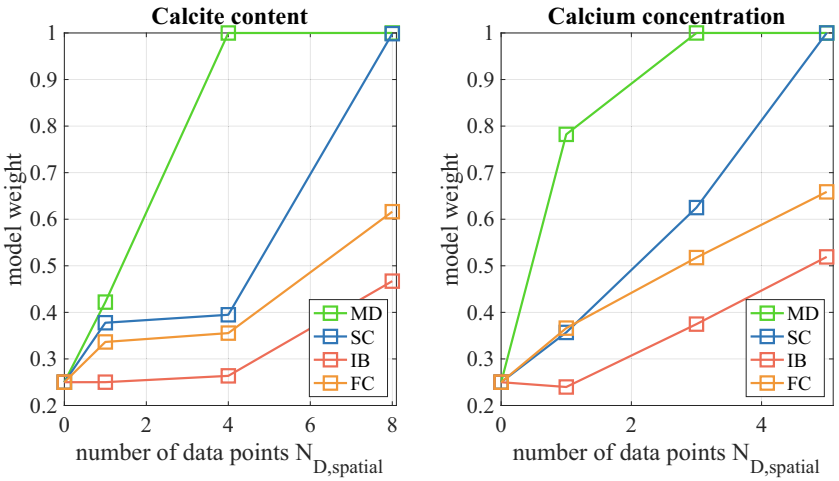


Fig. 7 Average model weights for the data-generating process of the two quantities of interest (calcite content and calcium concentration) of the three models and the measurement data (MD) over increasing amount of used spatial data points $N_{D,\text{spatial}}$

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}

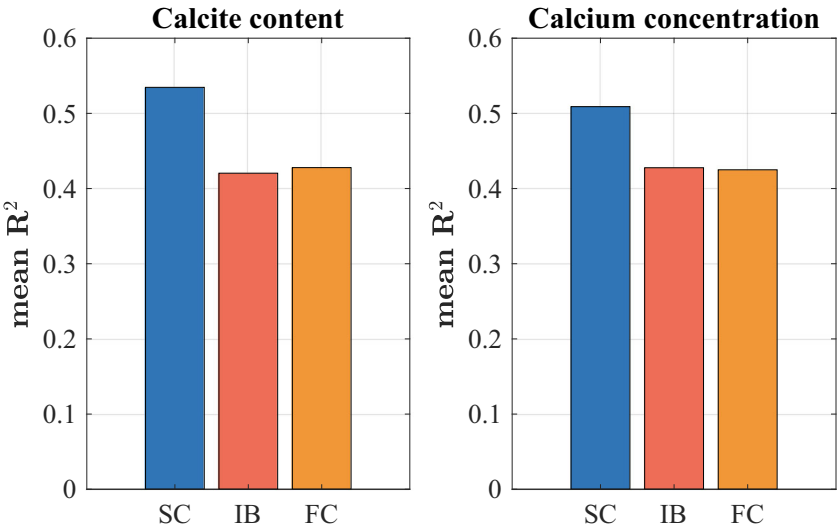


Fig. 8 Mean R^2 between the different model outputs and the measurement data

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

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

A.3 Diagnosing Similarities in Probabilistic Multi-Model Ensembles - an Application to Soil-Plant-Growth-Modeling

Modeling Earth Systems and Environment
https://doi.org/10.1007/s40808-022-01427-1

ORIGINAL ARTICLE

Diagnosing similarities in probabilistic multi-model ensembles: an application to soil–plant-growth-modeling

Aline Schäfer Rodrigues Silva[1] · Tobias K. D. Weber[2] · Sebastian Gayler[2] · Anneli Guthke[1,3] · Marvin Höge[1,4] · Wolfgang Nowak[1] · Thilo Streck[2]

Received: 10 December 2021 / Accepted: 9 May 2022

Abstract
There has been an increasing interest in using multi-model ensembles over the past decade. While it has been shown that ensembles often outperform individual models, there is still a lack of methods that guide the choice of the ensemble members. Previous studies found that model similarity is crucial for this choice. Therefore, we introduce a method that quantifies similarities between models based on so-called energy statistics. This method can also be used to assess the goodness-of-fit to noisy or deterministic measurements. To guide the interpretation of the results, we combine different visualization techniques, which reveal different insights and thereby support the model development. We demonstrate the proposed workflow on a case study of soil–plant-growth modeling, comparing three models from the Expert-N library. Results show that model similarity and goodness-of-fit vary depending on the quantity of interest. This confirms previous studies that found that "there is no single best model" and hence, combining several models into an ensemble can yield more robust results.

Keywords Multi-model ensembles · Energy statistics · Model set visualization · Crop modeling

Introduction

Multi-model approaches

Multi-model ensembles have received increasing interest in crop-modeling over the last decade (Palosuo et al. 2011; Asseng et al. 2013, 2015; Martre et al. 2015; Wöhling et al. 2015; Yun et al. 2017; Makowski 2017; Wallach et al. 2018). While multi-model approaches can serve different purposes (Höge et al. 2019; Minka 2002), the main focus in the crop-modeling community has been to improve the accuracy of predictions (Asseng et al. 2015; Martre et al. 2015) or to estimate the uncertainty due to model choice, often referred to as conceptual uncertainty (Asseng et al. 2013). Please note that no multi-model method can quantify the conceptual uncertainty on an absolute level (Nearing and Gupta 2018). The reason for this is evident: In practice, there is no way to create and sample from an exhaustive list of all plausible

✉ Aline Schäfer Rodrigues Silva
aline.schaefer@iws.uni-stuttgart.de

Tobias K. D. Weber
tobias.weber@uni-hohenheim.de

Sebastian Gayler
sebastian.gayler@uni-hohenheim.de

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

Marvin Höge
marvin.hoege@eawag.ch

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

Thilo Streck
thilo.streck@uni-hohenheim.de

[1] Department of Stochastic Simulation and Safety Research for Hydrosystems, Institute for Modelling Hydraulic and Environmental Systems/Cluster of Excellence "Data-Integrated Simulation Science", University of Stuttgart, Pfaffenwaldring 5a, 70569 Stuttgart, Germany

[2] Department of Biogeophysics, Institute of Soil Science and Land Evaluation, University of Hohenheim, Emil-Wolff-Straße 27, 70593 Stuttgart, Germany

[3] Junior Research Group for Statistical Model-Data Integration, Cluster of Excellence "Data-Integrated Simulation Science", University of Stuttgart, Pfaffenwaldring 5a, 70569 Stuttgart, Germany

[4] Department of Systems Analysis, Integrated Assessment and Modelling, Eawag-Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, Dübendorf 8600, Zurich, Switzerland

Published online: 19 June 2022

models that cover the entire range of possible outcomes (Nearing and Gupta 2018; Vehtari and Ojanen 2012; Höge et al. 2019; Ferré 2017).

Therefore, Nearing and Gupta (2018) suggest understanding multi-model methods rather as sensitivity analyses. From this point of view, multi-model methods are tools that make modelers aware of how much predictions may differ within the model set depending on the choice for a certain conceptual model. This follows the line of thought of Ferré (2017) who introduced the idea of a multi-model ensemble as a "team of rivals", which provides competing views of a system. If the competing models agree on a certain prediction, this increases the decision makers' confidence, while disagreement indicates the need for further investigation (Ferré 2017).

When different models are merged to *improve* predictions, modelers hope for two effects that make the ensemble more skillful than its individual members: (1) the errors of the individual models cancel one another. This requires the individual models to be *independent*, i.e. different in their assumptions and conceptualizations (e.g. Abramowitz and Gupta 2008; Abramowitz 2010; Evans et al. 2013; Sanderson et al. 2015a, b; Abramowitz et al. 2018; Enemark et al. 2019). However, this assumption is often not met (Abramowitz and Gupta 2008; Abramowitz 2010; Bishop and Abramowitz 2013; Evans et al. 2013; Sanderson et al. 2015a, b; Knutti et al. 2017; Abramowitz et al. 2018). (2) The ensemble covers a broad spectrum of possible system behavior (Enemark et al. 2019) and thus compensates for over-confident individual models (e.g. Fritsch 2000; Doblas-Reyes et al. 2005; Weigel et al. 2008).

Various studies have compared the predictive performance of multi-model ensembles to the ones of the individual ensemble members (e.g. Krishnamurti et al. 2000; Georgakakos et al. 2004; Doblas-Reyes et al. 2005; Hagedorn et al. 2005; Weigel et al. 2008; Diks and Vrugt 2010; Arsenault et al. 2015; Yun et al. 2017; Christiansen 2018). They found the following aspects to be crucial for the performance of ensemble predictions:

1. the applied combination/averaging/weighting method (Krishnamurti et al. 2000; Doblas-Reyes et al. 2005; Diks and Vrugt 2010; Arsenault et al. 2015; Höge et al. 2019, 2020),
2. the performance measure used for assessing the predictive skills (Hagedorn et al. 2005; Weigel et al. 2008; Diks and Vrugt 2010),
3. the spread (variability) of the individual models (Fritsch 2000; Doblas-Reyes et al. 2005; Weigel et al. 2008), and
4. the similarity of the individual models regarding their structure and their predictions (Tebaldi and Knutti 2007; Abramowitz and Gupta 2008; Abramowitz 2010; Winter and Nychka 2010; Bishop and Abramowitz 2013; Evans et al. 2013; Sanderson et al. 2015a, b; Abramowitz et al. 2018; Enemark et al. 2019).

In the present study, we focus on the last two of these aspects: the similarity and the spread of the individual models. We show how the similarities of (or distances between) the models can be quantified while accounting for the spread in their predictions.

Referring to the often claimed superiority of ensemble predictions, Hagedorn et al. (2005) argue that the question of whether the ensemble has higher predictive skill than the best individual model is posed wrongly because there is often no "best" individual model. They show that it is usually not the same model that performs best considering all quantities of interest or under all conditions. Rather, what is typically identified as the "best" model looking at a particular aspect of the simulated system might be a weak model considering another aspect. Therefore, Hagedorn et al. (2005) conclude that the ensemble is superior to single models because its predictions are more robust, i.e. better over a broad range of predicted variables and modeling periods. We set up the present analysis accordingly, such that it enables a model comparison considering various quantities of interest.

Ensemble predictions in crop modeling

In crop modeling, a systematic assessment of multi-model ensembles was initiated within the Agricultural Model Intercomparison and Improvement Project (AgMIP; Rosenzweig et al. 2013). In the wheat pilot study of that project, Asseng et al. (2013) compared 27 wheat models by analyzing the predicted grain yield under climate change conditions. They found that predictions vary significantly between different models. Thus, there is considerable uncertainty concerning model choice when predicting yields under climate change conditions. In an earlier comparison of yield predictions from eight crop models, Palosuo et al. (2011) also found great differences between the individual models. This study showed that none of the models was able to outperform its competitors across different environmental conditions and for different variables. In addition, the comparison of four crop models by Wöhling et al. (2015) had similar findings that support the argument of Hagedorn et al. (2005) that there is no "single best model".

While Asseng et al. (2013) focused on the "end-of-season variable" grain yield, Martre et al. (2015) compared the same 27 models also regarding grain protein concentration and "in-season variables" (leaf area index, plant-available soil water, total aboveground biomass, total above-ground nitrogen, and nitrogen nutrition index), with all models being calibrated on phenology data. The authors found that the ensemble predictions are more reliable and attributed this improvement to the different process descriptions providing

a wide range of plausible system behavior. The study also reports that some models had rather small errors for the end-of-season variables yield and grain protein concentration while showing large errors for in-season variables. Therefore, Martre et al. (2015) emphasize that for comparing the performance of different models, it is important to consider several variables as a model might perform well regarding a certain variable, but poorly regarding others.

Another suggestion of Martre et al. (2015) is to further investigate how to choose the individual models for an ensemble and how to weigh them. Many studies in the climate modeling community found that the dissimilarity of the individual models is crucial for the success of an ensemble (e.g. Tebaldi and Knutti 2007; Abramowitz and Gupta 2008; Abramowitz 2010; Sanderson et al. 2015a, b; Abramowitz et al. 2018). If two models in the ensemble are highly similar, this leads to difficulties in the weighting scheme as these models should not receive the same weight as if they were independent. George (2010) and Garthwaite and Mubwandarikwa (2010) therefore recommend using so-called "dilution priors" that divide the weight between partly redundant models. We hypothesize that quantifying the similarity between the models can help to choose the individual ensemble members and weigh them in a way that accounts for possible redundancies.

Another ubiquitous issue when comparing models is calibration. In a recent study, Wallach et al. (2020) discuss the "chaos in calibrating crop models", i.e. the lack of a unified calibration procedure in the crop modeling community. In an earlier study, Wallach (2011) stated that in the calibration of crop models, often model structural errors are compensated by specifying non-physical parameter values. As a result, a model might perform well for the quantity of interest it has been calibrated on, but poorly for others. This effect is more severe, the more parameters a model has (e.g. Jefferys and Berger 1992; Lever et al. 2016). Therefore, Vogel and Sankarasubramanian (2003) recommend checking model adequacy in an uncalibrated state. We follow that recommendation and implement our analysis in a Monte Carlo framework, sampling prior parameter distributions. This allows us to evaluate the model performance independent of a specific parameter choice and we avoid to cloud model errors by assigning parameter values that compensate for structural deficiencies. As in any Bayesian framework, a subjective choice of prior distributions based on expert knowledge is needed. In fact, in the Bayesian setting, plausible ranges and assumed distributions are part of the model, just as a fixed parameter assumption (or the decision that a parameter is free for calibration) would be part of a model in deterministic modelling. Different choices among these options makeup different models.

Goal and approach of this study

The main goal of this study is to quantify similarities between probabilistic model predictions and visualize them intuitively. The proposed methods help modelers to gain deeper insights into the model set and to choose a suitable multi-model strategy accordingly. Our approach is to use a statistical metric, the so-called energy distance (Rizzo and Székely 2016) to quantify the (dis-)similarity between probabilistic model predictions and noisy measurements. Metrics, in general, are distance measures. Statistical metrics measure how close statistical objects such as probability distributions are, i.e. they take probability densities into account. We use the energy distance as a metric to compare predictive distributions that are generated by sampling from the prior parameter distributions of each model in a Monte Carlo framework. This enables us to take parametric uncertainty into account and to compare the models independent of a specific parameter choice. Thus, for comparing the models, we calculate the energy distance between their predictive distributions.

With the same method, we can also assess model performance by calculating the energy distance between model predictions and noisy observations. For this, we fit a probability density function to replicate measurements. From this distribution, we draw samples to calculate the energy distance in a Monte Carlo framework. If no distribution for the measurement errors can be defined (e.g. because no replicate measurements are available and no assumptions about measurement noise seem defendable), we can use the deterministic counterpart of the energy distance: The so-called energy score (Gneiting and Raftery 2007) compares probabilistic distributions to deterministic values and is directly related to the energy distance. This makes the concept of energy statistics widely applicable for rating probabilistic models.

The proposed method fulfills the following properties:

1. It can act on multivariate model predictions and thus reflect "overall" model characteristics.
2. It quantifies the similarity between pairs of models in the same way as the similarity between models and observations. Thus, it can assess the similarity between model predictions as well as model performance given observed data.
3. It can be used for comparing probabilistic model predictions to both noisy observations (by using the energy distance) and to deterministic observations (by using the energy score).
4. It acts on prior predictive distributions and thus accounts for parametric uncertainty in each model.

For guiding the multi-model process, we need an intuitive way to visualize the quantified similarities among the

models and measurements. Therefore, we suggest different methods for visualizing similarities among models, which highlight different aspects of similarity and, when combined, provide a detailed overview for interpreting the model set.

The paper is structured as follows: First, we present the mathematical methods, i.e. the "Energy distance" and the "Energy score", and visualization techniques in the section "Visualizing predictive similarity". Experimental data are presented in "Field experiments" and we introduce the model set in "Model description". This is followed by "Results and discussion". We summarize our findings and provide conclusions in the section "Summary and conclusions".

Methods

Energy distance

In this section, we describe how the (dis-)similarity between two probabilistic models or between a probabilistic model and noisy measurements can be quantified and visualized based on Monte Carlo samples of the models' predictive distributions.

Well-known distance measures like the Euclidean or Manhattan distance (L_p-metrics) are based on the coordinates of points in the Euclidean space. These distances do not take the density of probability distributions into account. In contrast, statistical distances (also known as probability metrics) measure the distance between two statistical objects such as random variables, probability distributions, or data samples (e.g. Deza and Deza 2016) and include information about probability densities.

Rizzo and Székely (2016) introduced the energy distance as a metric that measures the distance between two random vectors X, Y in $\mathbb{R}^D$. It is called energy distance because of the analogy to the potential energy between objects (Rizzo and Székely 2016). It satisfies all axioms of a distance metric (non-negativity, identity of indiscernibles, triangle inequality) (e.g. Deza and Deza 2016). The squared energy distance D^2 between the distributions $F(X)$ and $G(Y)$ is defined as

$$D^2(F,G) = 2\mathbb{E}||X-Y||_2 - \mathbb{E}||X-X'||_2 - \mathbb{E}||Y-Y'||_2 \geq 0, \tag{1}$$

with $\mathbb{E}$ being the expected value, $||\cdot||_2$ being the Euclidean norm, X and X' being independent and identically distributed (iid) variables, the same applies for Y and Y'. In this study, we analyze data based on the energy distance $d(F,G) = \sqrt{D^2(F,G)}$.

The expected values in Eq. 1 can be implemented in a Monte Carlo framework as follows:

$$\mathbb{E}||X-Y||_2 = \frac{1}{N_{\mathrm{MC}}^2}\sum_{k=1}^{N_{\mathrm{MC}}}\sum_{l=1}^{N_{\mathrm{MC}}}\sqrt{\left(\mathbf{x}_k - \mathbf{y}_l\right)^2}, \tag{2}$$

where $x \sim F$, $y \sim G$, and N_{MC} being the number of Monte Carlo samples.

Figure 1 shows four 1D examples that illustrate how the energy distance between two univariate probability density functions (pdf) changes depending on the mean Euclidean distance between these pdfs $\mathbb{E}||X-Y||_2$ and the mean Euclidean distance within each pdf $\mathbb{E}||X-X'||_2$. Analogously, the energy distance can quantify the distance between D-dimensional random vectors.

Comparing Fig. 1a, d as well as (b) and (e) shows that keeping the same mean and increasing the variance of distribution G decreases the energy distance $d(F, G)$ between both distributions. Subfigure (c) shows that for two identical distributions, the energy distance becomes zero, while the expected value of the Euclidean norm $\mathbb{E}||X-Y||_2$ is not equal to 0. Subfigure (f) illustrates the energy distance between two distributions with the same mean but different variances.

Energy score

When working with real and error-prone data, we do not have access to the full distribution of the data (i.e. a "true" value and a distribution function of errors) but only to the measured instances thereof, i.e. our observations. In some cases, these measurements suffice for estimating the underlying distribution reasonably well. If this is not the case (e.g. if there are only a few measurements available), we need an alternative for rating probabilistic predictions given deterministic measurements.

In deterministic modeling, the performance of a model is usually evaluated by an error measure between the model's best estimate $\hat{\mathbf{y}}_k$ and the observations $\mathbf{y}_{\mathrm{meas}}$. Different models are then rated based on the achieved best estimate error, e.g. a root mean square error (RMSE) or the mean absolute error (MAE).

In probabilistic modeling, model rating is based on so-called scoring rules (Gneiting and Raftery 2007). These scores account for the entire predictive distribution of the model instead of only the best estimate. Many different scores exist (Gneiting and Raftery 2007; Yao et al. 2018). Among these, the energy score is directly related to the above-introduced energy distance (Székely and Rizzo 2013), i.e. it resembles the one-sided version of the energy distance (Ziel and Berk 2019). The energy score ES for the model predictive distribution G and observations $\mathbf{y}_{\mathrm{meas}}$ writes as:

$$\mathrm{ES}(G, \mathbf{y}_{\mathrm{meas}}) = \frac{1}{2}\mathbb{E}||Y-Y'||_2^{\beta} - \mathbb{E}||Y-\mathbf{y}_{\mathrm{meas}}||_2^{\beta}, \tag{3}$$

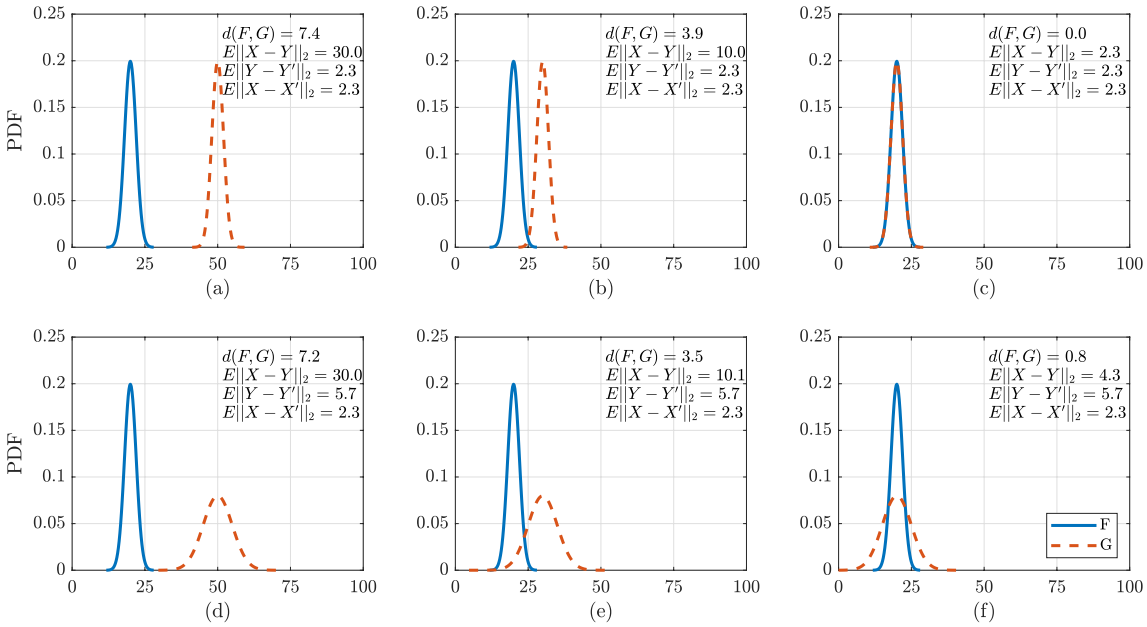


Fig. 1 Illustrative 1D example of the energy distance $d(F, G)$ between two probability density functions F and G. It is calculated based on the mean Euclidean distance between these functions $\mathbb{E}||X - Y||_2$ and the mean Euclidean distance within each function $\mathbb{E}||X - X'||_2$ and $\mathbb{E}||Y - Y'||_2$, respectively

with $\beta \in (0, 2)$. We choose $\beta = 1$ as it is a standard choice for distributions that are not heavily tailed (Ziel and Berk 2019). For $\beta = 2$, the energy score is equal to the negative squared error (Gneiting and Raftery 2007).

In cases we cannot assume a reasonable distribution based on replicate measurements, we will use the energy score instead of energy distance. We want both quantities to act on the same scale, so they are directly comparable. Therefore, we use $d(F, G) = \sqrt{D^2(F, G)}$ and $\sqrt{\text{ES}}$ for the analysis.

Visualizing predictive similarity

We want to visualize the (dis-)similarity of the model predictions to get an intuitive understanding of the diversity in the considered model set. At the same time, we want to visualize how well the predictions match the measurements. Therefore, we treat both the models and the measurements as objects in a common model predictions-observations space, which we call "quantity of interest space". Representing the similarities of N objects (models and observed data) leads to $n_{\text{comb}} = N \cdot (N - 1)/2$ combinations. While in our application, this number (three models and one measurement data set, hence, six combinations) is comparatively small, visualization of model similarity in two dimensions is already not a straightforward task. Clearly, the number of models to be compared can become much higher in extensive multi-model ensembles. Therefore, the methods we propose for visualization are also suitable for larger model sets.

Each of these objects (models and observed data) consists of n_{qoi} variables (quantities of interest). In the case of probabilistic modeling, each variable is assigned a probability distribution. Therefore, we have to deal with high-dimensional data, and regarding its two-dimensional visualization, we have to balance the interpretability and the preservation of the original structure in the applied projection (Liu et al. 2017), which is a typical problem in the visualization of high-dimensional data.

We make use of different techniques for visualizing the similarity of two objects (model-model, model-data) under different conditions. Each visualization method highlights a different aspect, so which method is the most insightful one depends on the specific question we ask about the model set:

1. Heatmaps: In a matrix, the distances between all pairs of objects are visualized through varying colors or intensity (e.g. Nandi and Sharma 2020).
2. Radar charts: Several axes are arranged radially starting from a common center. Each axis represents a certain quantity of interest, i.e. a different variable or the same variable under different boundary conditions. Each value (here, the distance between two objects) is plotted along one axis. This is repeated for all axes. Finally, all values are connected to a polygon, representing one object (e.g. Nandi and Sharma 2020).
3. Dendrograms: Dendrograms are tree-like diagrams that are typically used for visualizing hierarchical structures. A dendrogram consists of branches that connect objects

depending on their similarity. The height at which two objects are joined together represents the distance between these objects (e.g. Nandi and Sharma 2020). For creating such a diagram, we use an agglomerative hierarchical clustering approach (Xu and Wunsch 2008): An algorithm identifies pairs of clusters with minimal distance in-between and merges them. This merging procedure is repeated until all data points are finally in one overarching group (Xu and Wunsch 2008). The merging depends on the chosen linkage method, i.e. the definition of the distance between two clusters. For the present study, we chose a linkage that uses the average distance between data points in two clusters.

Case study description

Central to our study is the simulation of wheat growth, energy and water fluxes in six agricultural fields in two regions during several years (2010–2015). These fields feature slightly contrasting meteorological conditions at two different sets of sites (1–3 and 4–6), with soils being only similar in sites 1–3. The study is based on an extensive data set enabling coupled soil–plant-growth modeling and comparing simulation results to high-quality measurements. The field data is summarized in "Field experiments" and the participating models of the ensemble in "Model description".

Field experiments

The data set used in this study is a subset of a multi-site, multi-year, and multi-crop data set that contains extensive characterization of soil properties and states, plant growth and yield, management, and soil–atmosphere fluxes of energy, water and carbon dioxide. It was obtained from measurement campaigns in intensively managed agricultural fields of local farmers. We will limit the description of the data set to a minimum since it was published alongside a manuscript with full methodological details (Weber et al. 2021).

The data were collected between May 2009 and September 2018. In this study, we use a subset that covers the sites and years in which winter wheat was cultivated from 2010–2015 (year of harvest). The combination of a site at which and a year during which winter wheat was cultivated are reported as site-years. For example, winter wheat grown at site 1 from November 2014 to July 2015 is denoted as site 1, year 2015. In total, we analyze data of 14 site-years. Details of the two regions can be found in Weber et al. (2021), and their soil properties are summarized in Table 1.

All required meteorological forcings were measured at half-hourly time intervals and data gaps were filled. In the data set, grain yield was reported both by the farmer as a field average and by extrapolation of the plot sampling to the field. Phenology and leaf area index were measured at least biweekly during the main vegetation season (April to mid July).

Table 1 Properties of the soil horizons at sites 1–6, and the soil hydraulic property model parameters θ_s and θ_r, which are the fixed saturated and residual water content, respectively

Site	Depth	No. of simulation layers	Organic matter	Clay	Silt	Sand	Bulk density	θ_r	θ_s
(–)	(cm)	(–)	($gg^{-1} \cdot 100$)				($g\,cm^{-3}$)	($cm^3\,cm^{-3} \cdot 100$)	
1	0–30	6	1.75	18.2	79.4	2.5	1.37	7.5	46.0
	30–45	3	0.61	18.8	79.2	2.0	1.51	7.2	42.9
	45–165	24	0.42	18.7	80.4	0.9	1.48	7.3	43.9
2	0–30	6	1.53	17.9	79.5	2.6	1.33	7.6	47.0
	30–45	3	0.52	20.1	77.0	2.9	1.46	7.5	43.9
	45–165	24	0.34	18.7	79.7	1.6	1.53	7.1	42.5
3	0–30	6	1.64	17.1	81.1	1.8	1.37	7.4	46.2
	30–45	3	0.83	18.7	80.4	1.0	1.50	7.3	43.4
	45–165	24	0.63	16.1	83.0	0.8	1.51	6.9	43.2
4	0–20	4	4.35	37.8	56.0	6.2	1.31	9.5	49.2
	20–30	2	2.13	38.6	52.5	8.9	1.34	9.4	48.2
	30–50	4	1.63	48.4	43.3	8.4	1.32	10.0	50.0
5	0–20	4	3.64	28.9	68.3	2.8	1.37	8.6	46.8
	20–60	8	1.44	33.6	64.3	2.1	1.40	8.9	46.7
	60–165	21	0.71	34.2	64.0	1.8	1.51	8.7	44.0
6	0–30	6	5.50	45.6	51.2	3.2	1.04	10.7	59.0
	30–45	3	3.88	47.6	48.3	4.1	1.29	10.1	51.3

Model description

The relevant processes for crop development and growth, unsaturated water flow, nitrogen and carbon turnover in the soil, evapotranspiration, and drainage water quantity and quality were simulated with the multi-model library Expert-N (Priesack 2006). Expert-N is a model system that facilitates a high degree of flexibility in selecting competing model formulations for the relevant processes in the soil-vegetation-atmosphere continuum. An example of a 48 member multi-model ensemble using Expert-N is the study about climate change impact on wheat and maize yield development in Ethiopia by Rettie et al. (2022). That study uses a model ensemble consisting of 48 unique model members set up in Expert-N. For the presented study we selected three different plant growth models within Expert-N: CERES (Ritchie et al. 1988), SUCROS (van Laar et al. 1997), and SPASS (Wang and Engel 2000; Gayler et al. 2002), which are coupled to the soil carbon and nitrogen turnover and transport models SOILN (Johnsson et al. 1987) and LEACHN (Hutson and Wagenet 1995), and the Richardson–Richards equation for variably saturated water flow as implemented in HYDRUS-1D (Šmunek et al. 1998). Fluxes of heat and dissolved nitrogen in the soil were described by LEACHN and potential evapotranspiration as calculated by the Penman–Monteith equation (Allen 1998) modulated by crop coefficients. These models are described in "Phenological development"– "Soil nitrogen". The model initial and boundary conditions, including a description of the uniformly distributed and bounded model parameter priors for sampling the prior predictive distribution, are given in "Process models". In the following, SUCROS, CERES, and SPASS are described and refer to the versions implemented in Expert-N v.3.1. Model parameters and priors are listed in Tables 5, 6, and 7.

Phenological development

In Expert-N, the phenological development (BBCH) is modeled as a parametric function of thermal time, vernalization, photoperiod effect, and temperature sensitivity. While CERES is differentiated into nine development phases, SPASS and SUCROS are differentiated into three. All three adopted models distinguish the vegetative growth phase (BBCH ≤ 60) from the generative phase (BBCH > 60) and one for emergence. CERES (Ritchie and Godwin 1989; Jones 1986) and SUCROS (Spitters et al. 1989; Van Laar et al. 1992) are widely established models and SPASS is a combination and development out of the other two (Wang 1997; Gayler et al. 2002). Internally, phenological development is first simulated as a development stage on a scale from − 0.5 to 2.0, and is subsequently converted to an externally reported BBCH variable using fixed lookup tables with 10 support points (11 in the case for SPASS). The support points in the lookup tables were considered as fixed. The simulated phenology acts as a boundary condition for the remaining part of the dynamic plant growth model, by setting the precondition after which certain other parts of the model are active (i.e. triggering submodules for, e.g. leaf area index or grain filling after anthesis at BBCH = 60). An important difference between the models is that winter wheat requires vernalization, which is the induction of flowering after a cold period. In contrast to CERES and SPASS, which contain routines for vernalization. Since SUCROS does not include vernalization, it is strictly speaking, a spring wheat model.

Root growth, root water uptake, and transpiration

Dynamic root growth in all plant models is simulated by roots growing downward up to a maximum root extension depth. The maximum growth rate is reached under optimum conditions. This is modulated by impacts of unfavorable environmental factors (temperature, soil moisture) in the layer of the currently greatest root extension. Specifically, these impedances are functions of temperature and low soil moisture (SPASS, SUCROS), or of low soil moisture and low mineral nitrogen contents (CERES). In each simulation layer, the active roots are the balance between root growth and senescence at each time step (Gayler et al. 2013). To calculate root water uptake, a root length density is required. CERES and SPASS use an identical approach: the root length growth rate is linearly related to the root biomass growth rate, and the vertical distribution is related to water and nitrogen availability in the respective soil layers. In SUCROS, the root length growth rate is derived for each simulation layer based on a crop-specific root depth distribution function, the root biomass, and a specific root length. The upper limit of root water uptake is limited to the potential transpiration as calculated by the Penman–Monteith equation (Allen 1998). In the case of SPASS and CERES, a maximum root water uptake rate per root length is additionally defined. All three models use a macroscopic approach in which root water extraction is distributed to the individual simulation layers proportionally to the relative root length in the layers, as long as the water supply is optimum. Impedance factors such as oxygen deficiency in (near-)saturated soils, soil compaction and structure, disease and pests, adverse chemical conditions (e.g. salts) are not considered in the models. To account for crop effects on potential evapotranspiration, all three models use crop coefficients, which we modeled as a piece-wise phenology-state dependent function with three parameters kc_{ini} (−), kc_{mid} (−), and kc_{end} (−), which are the crop coefficient for the initial, mid and end of the vegetation period, respectively. These

parameters were considered uncertain with the uniform distribution in the ranges given in the Appendix, Table 4.

CO_2 assimilation, biomass growth and leaf area development

CERES adopts a robust "big leaf" approach to calculate carbon assimilation, using empirical adjustments to account for the depth-dependence of photosynthetic capacity and light response, with the interception of photosynthetically active radiation dependent on leaf area index. Biomass development depends on the partitioning of assimilates to different plant organs (roots, leaves, stem, and fruit). It is achieved through potential growth rates and a priority scheme for the allocation of assimilates to each organ (differentiated into five developmental stadia). In the juvenile phase, i.e. in the stadium between the emergence and, the development of the first apical spikelet, the leaf area index develops exponentially, and after the juvenile phase, leaf area develops proportionally to the leaf biomass development, depending on temperature, and water and nitrogen stress. In SUCROS and SPASS, the calculation of carbon assimilation is based on a multi-layer approach, which is more comprehensive compared to CERES. The aim of this approach is to differentiate between sunlit and shaded leaves and to account for the attenuation of direct and diffuse radiation. The two models differ in vertical resolution as SUCROS uses a three-layer approach, and SPASS uses five layers, but the calculation of leaf internal CO2 concentration and net photosynthesis is similar, with small differences in the calculation of water stress and nitrogen response functions. In contrast to CERES, in SPASS and SUCROS carbohydrate allocation and hence organ growth follows an assimilate-partitioning scheme, which is fixed at optimum water supply and is determined solely by the development stage of the plant. However, in the case of water deficiency, root growth is favored in both models to counteract the cause of stress. In SUCROS, the leaf area growth is directly coupled to leaf biomass growth rate, whereas in SPASS, leaf area growth rate does account for water and nitrogen deficiency. More detailed presentations can be found in Priesack (2006), Priesack and Gayler (2009), Biernath et al. (2011), Wöhling et al. (2013) and references therein.

Soil hydrology

In Expert-N, the standard process model for simulating variably saturated moisture fluxes in soils is the Richardson–Richards equation (RRE) (Richardson 1922; Richards 1931). The solution of the RRE requires parameter functions to describe the soil hydraulic properties. Since we simulated root water uptake using a macroscopic approach (van Dam et al. 2008), it is sufficient to parameterize the RRE using the van Genuchten Mualem model (van Genuchten 1980). There would be physically more comprehensive soil hydraulic property models that account for water storage and conductivity in medium to dry soils (Weber et al. 2019; Streck and Weber 2020; Weber et al. 2020). These would influence the simulation of actual transpiration under water-stressed conditions only when using microscopic (not macroscopic) root water uptake models (van Dam et al. 2008), providing hydraulic uplift does not influence the simulation.

For each site, a top-soil/sub-soil differentiation was made, each with a different parameterization. The varied parameters per soil layer are α (cm^{-1}) and n (–), which are the van Genuchten shape parameters, the saturated hydraulic conductivity K_s ($cm\,d^{-1}$), and the tortuosity parameter τ (–), and we fixed $m = 1 - 1/n$. Instead of varying the saturated and residual water contents, $\theta_r(-)$ and θ_s, respectively, we vary the soil water content profile set as an initial condition. The soil model is discretized into simulation layers of 5 cm depth (see Appendix, Table 1). Here, in contrast to the observed soil profiles, we reduce the number of simulated soil horizons for the soil hydrological part to two by merging the second and third horizon at sites 1–5. The differentiation into more horizons is pedologically founded, but for the modeling purpose of this study not parsimonious, i.e. we group horizons with very similar hydraulic properties.

Soil nitrogen

Mineralization, nitrification, and denitrification are modeled following the SOILN approach, while urea hydrolysis, volatilization and dissolved nitrogen transport are modeled using the LEACHN approach.

The model concept of SOILN differentiates three pools of organic nitrogen representing the three different pools of organic carbon, available to the soil microbes. These pools are termed 'litter', 'humus', and 'manure'. The 'litter' pool, a pool with fast turnover rates, represents fresh organic matter and microbial biomass. The 'humus' pool with a slow turnover of soil organic carbon, and the 'manure' pool, which represents the organic fertilizer. There are two essential assumptions of this model concept; (i) the N demand for the internal carbon cycle is governed by a constant C/N ratio of 10:1 in the microbial biomass and in the humus pool, and (ii) mineral nitrogen that is released or assimilated by the microbial biomass, follows this ratio. We varied the rate constants $k_{\mathrm{miner,man}}$ (d^{-1}), $k_{\mathrm{miner,lit}}$ (d^{-1}), $k_{\mathrm{miner,hum}}$ (d^{-1}), which govern the rate of mineralization of the manure, litter, and humus pools. We also varied the rate coefficients for the nitrification k_{nit} (d^{-1}) and denitrification $k_{\mathrm{miner,lit}}$ (d^{-1}), where, analogously to the treatment of the soil hydraulic properties, we model the two horizons with different sets of kinetic rate constant parameters, except for $k_{\mathrm{miner,man}}$, which we set to 0 for the sub-soil. The rate constant for the urea

hydrolysis $k_{urea,hy}$ is a constant 0.36 (d^{-1}). The effectivity of decomposition f_e (–) describes the fraction of carbon that is re-immobilised after decomposition and set to a constant 0.45. The humus development constant f_h (–) describes the fraction of decomposed litter that is added to the humus pool and was set to 0.2.

Model setup and implementation

Process models

As an upper boundary, we generally use in situ measured daily aggregated atmospheric temperatures (minimum, mean, and maximum), global radiation, wind speed, and precipitation. Potential evapotranspiration fluxes were calculated based on the Penman–Monteith approach. For the solutes, we use flux boundary conditions prescribing constant atmospheric NH4 deposition, and the timing, frequency, type, and amount of fertilizers. At the lower boundary, we used free drainage for the water flow module and a zero gradient for the solute and heat flux modules. Field management in terms of nitrogen input by fertilizers, and sowing and harvest dates were set to the farmer-reported data.

The selection of which parameters are considered as uncertain priors, and their respective ranges and distributional assumptions was guided by both model system and expert knowledge. Details are listed in Tables 4, 5, 6 and 7). We ran simulations based on $n_{MC} = 10{,}000$ parameter vector realizations generated by Latin Hypercube sampling per site-year and model. The gained forward simulation results resemble approximations to the prior predictive distributions for each of the 14 site-years and each of the plant models. This resulted in a total of 420, 000 individual simulations performed on the High-performance Cluster bwFOR of the Federal State of Baden-Württemberg. The varied model parameters are listed in the Appendix in Tables 4, 5, 6 and 7.

Similarity analysis

We analyze the similarity of probabilistic predictions of CERES, SUCROS, and SPASS (Priesack 2006) via the energy distance between the predictive distributions of the models for different variables. The analyzed variables are yield, phenology, and leaf area index. For comparing model predictions and measurements, we use either the energy distance or the energy score as discussed in "Energy distance" and "Energy score":

1. *Fitting a distribution to the observations and using the energy distance:* In the case of yield predictions, we can reasonably assume a Gaussian measurement error and, hence, define a distribution for the observations. Therefore, we can use the *energy distance* not only for the pairwise comparison of the models' distributions among each other but also for the comparison with the distribution fitted to the observations.
2. *Using the median of the observations and the energy score*: In the case of the other two variables, leaf area index and phenology, making assumptions about measurement error and fitting a distribution to the observations is not as straightforward. Instead, we take the median of the measurements and use the *energy score* to compare the models and the observations. For the comparison of the models among each other, however, we still use the *energy distance* because the predictive distributions are available.

Please remember that similarities quantified by the energy score and the energy distance are on the same scale, and hence directly comparable.

We analyze data from six sites and up to three years per site to check how model performance and similarity vary under different conditions.

Results and discussion

First, we compare the models' similarity (among each other) and their performance (i.e. similarity to observations) based on the end-of-season variable yield. Later, we analyze in-season variables to gain more insight into the processes that may have led to differences in the final yield predictions.

Analysis of the end-of-season variable yield

For yield, we assume a Gaussian measurement error, describing the distributions based on the replicates' mean and variance. The resulting distributions are shown in Figs. 2 and 3. Based on Monte Carlo samples of these distributions, we calculate the energy distance between models and observations.

Predictive distributions

Figures 2 and 3 depict the distributions for yield predicted by the three models and the measurements for the sites 1–3 and 4–6, respectively. The probability density functions represent the prior model predictions, i.e. the models have not been calibrated and represent the full range of plausible parameters as defined in Tables 5, 6 and 7.

The mean observed values for yield range from 7.0 t/ha (site 2, 2013) to 9.2 t/ha (site 4, 2014), the corresponding standard deviations range from 0.39 t/ha ≈ 5% (site 2, year 2011) to 2.1 t/ha ≈ 23% (site 6, year 2014).

The probability density functions show that the predictions made by SPASS (yellow) have the highest variance for

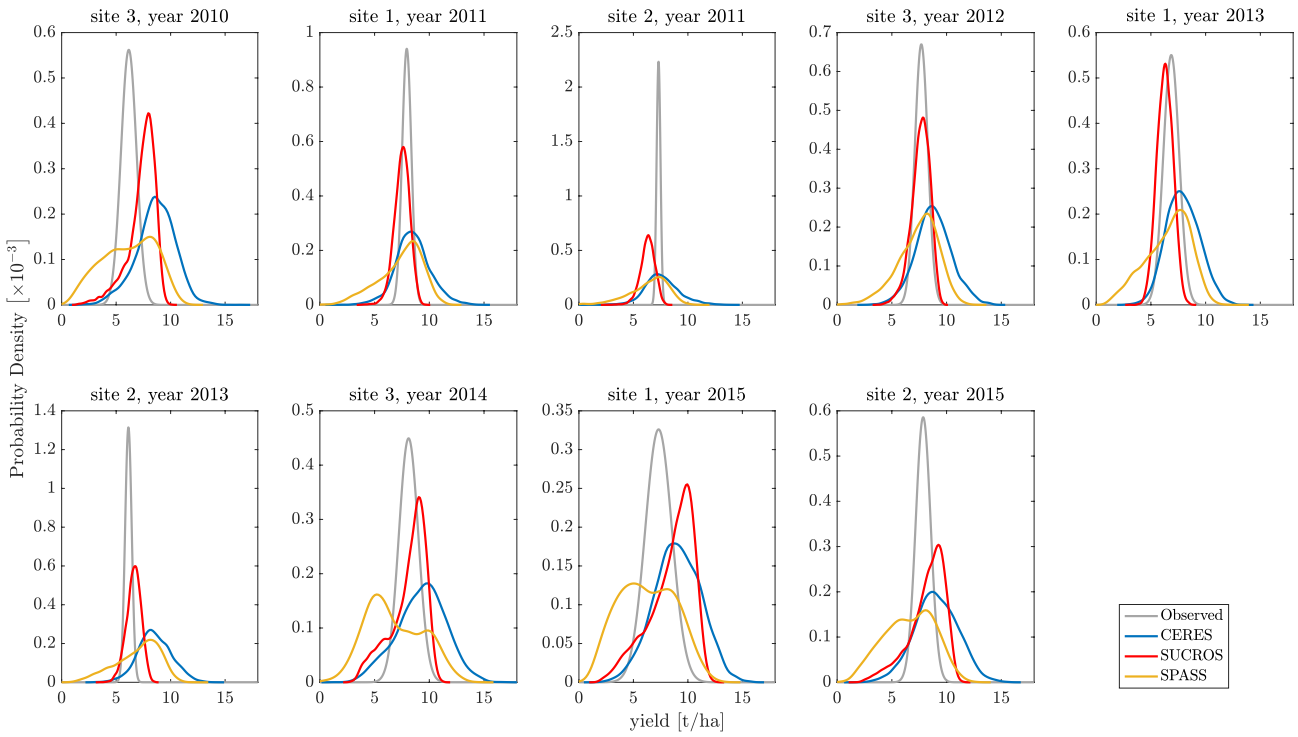


Fig. 2 Probability density functions of the yield predicted by the three models and observed (gray), sites 1–3. For better visualization, *y*-axis scales are not the same across all sub-plots

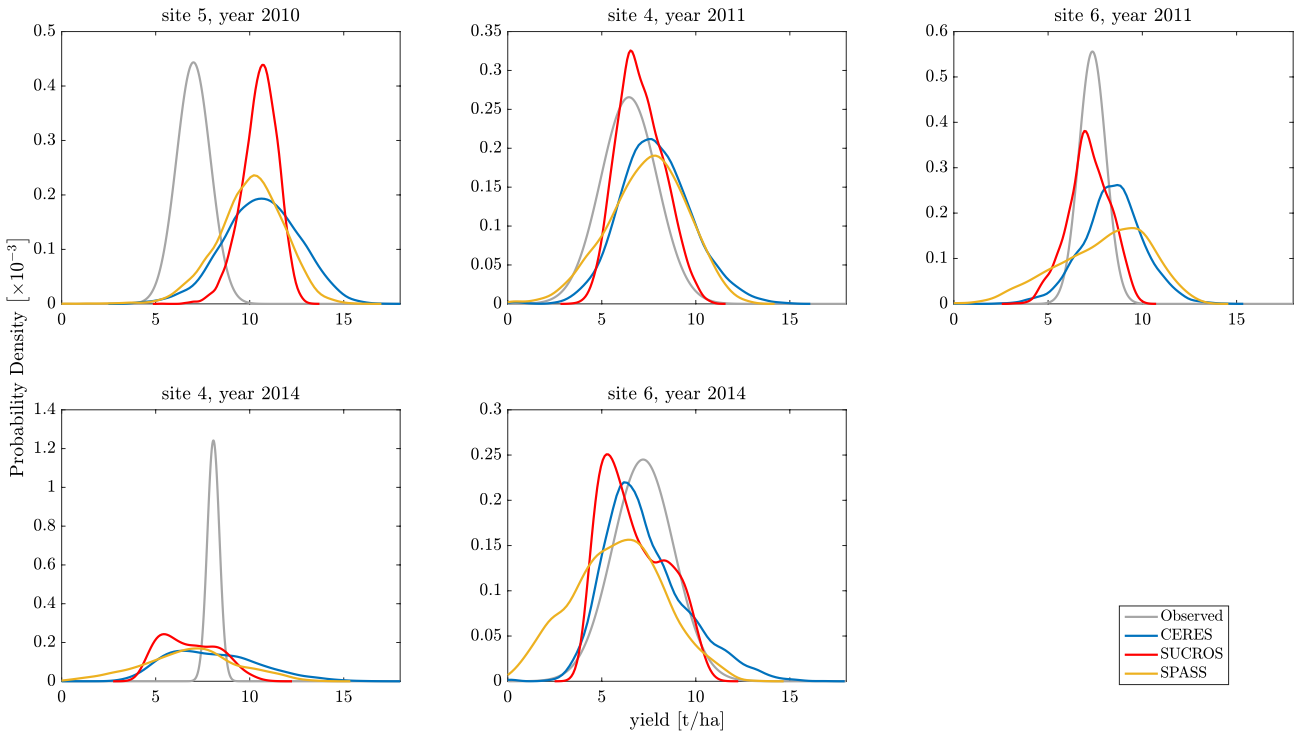


Fig. 3 Probability density functions of the yield predicted by the three models and observed (gray), sites 4–6. For better visualization, *y*-axis scales are not the same across all sub-plots

all sites and years and it predicts ranges of very low yield with a higher probability than CERES (blue) and SUCROS (red) do. For most cases, SUCROS shows the smallest variance in the yield predictions and, from the visual impression, the best goodness-of-fit to the measurements (gray).

Energy distance-based similarity analysis

To get a more aggregated and objective comparison of the predictive distributions of all models and the data, we quantify their similarity according to their prior predictive distributions using the energy distance (Eq. 1). In the following, we discuss corresponding visualizations with radar charts, dendrograms, and heat maps.

Visualizing model similarities using radar charts

Figure 4 shows four radar diagrams that represent the similarity of the models and the observations, each centered on one of the models, or the observations, respectively. In each chart, each model is represented by points that are connected across all radial axes. Each axis represents one site-year. The closer a point to the center, the lower the energy distance between the respective distributions, i.e. the more similar the distributions.

The background color of each segment is color-coded according to the weather conditions of the respective site-year. For this color-coding, we calculated the ratio of mean precipitation and mean temperature from April to June p/T and represent low values as red and high values as blue. This color-code is useful to investigate if there is any obvious relationship between weather conditions and model performance or similarity.

From Fig. 4a we can see that SUCROS is closest to the observed data for most site-years. In parts, this can be explained by SUCROS showing the lowest variance in the yield predictions, while being reasonably centered on the observations (see Figs. 2, 3).

Comparing model performances for sites 1 and 2 in 2013, we notice that CERES and SPASS are much closer to the data for site 1 than they are for site 2. Considering that the conditions at both sites are very similar, it may be surprising that the models perform so differently. The model predictions for both sites are indeed highly similar. However, the observations' variances differ considerably between both sites (see Fig. 2). Please recall the property of the energy distance that decreasing the variance of a distribution while keeping the same mean increases the energy distance between two distributions (see Fig. 1). This effect is clearly visible in the case of model performances for site 1 and 2 in 2013.

Focusing on the weather conditions during the growing season, we see that all models perform relatively poorly for site 5, 2010, which was a rather wet year at this site. Here, SPASS is slightly closer to the measurements than the other two models. This is due to the high variance of its predictions (see Fig. 2), which leads to a larger overlap of the predictive distribution with the measurement distribution, even though the modes of all three predictive model distributions are relatively similar, all overestimating the yield. For the other wet year 2013, we cannot observe a similarly poor performance of all models. Only the CERES predictions of site 2, 2013 show a rather high distance to the observations again overestimating the measured yield (see Fig. 2). However, for the relatively dry conditions at site 3 in 2010, CERES shows a similar distance as it again overestimates yield (see Fig. 2). Hence, in the current data set, the poor performance of the models cannot be clearly explained by specific weather conditions.

Next, we check whether any of the models perform above its average during certain weather conditions: CERES performs well for site 6 in 2014, which was a rather dry year. SUCROS shows equally good performance under both dry and wet years. SPASS performs comparatively well for site 4 in 2011, which was the second driest year in our data set. However, this is not true for other dry years in this data set.

In summary, we cannot observe a specific pattern of good or poor model performance under certain weather conditions. This result indicates that the data do not provide evidence for identifying systematic mispredictions based on climate or site. An exception is site 5, which, however, only has one replicate year, so that no general statement can be made.

In Fig. 4b–d, we focus on the similarity between the models. When CERES is in the center of the radar chart (subfigure (b)), we can see the high discrepancy between CERES and SPASS for site 3, 2014, site 1 and 2, 2015, site 3, 2010 (rather dry years). The highest distance between CERES and SUCROS occurs for site 1 and 2, 2013 (wet years). When SUCROS is centered (subfigure (c)), it is apparent that SPASS is closer, i.e. more similar for most site-years. Subfigure (d), with SPASS being centered does not contain new information that has not yet been shown in figures (a)–(c) and is only shown for the sake of completeness.

Visualizing model similarities using dendrograms

In Fig. 5, an alternative visualization of model similarities based on dendrograms is presented. Here, the data of sites 1–3 are shown, the corresponding Figure for sites 4–6 can be found in the appendix (Fig. 11). Please note that the order of the models is not necessarily the same for all dendrograms. Rather, this so-called leaf order was optimized such that the sum of the similarities between adjacent leaves is maximized (Novoselova et al. 2015).

From the way models and observations are merged into clusters and from the height at which two objects are joined together in the dendrograms, we can intuitively see their similarity.

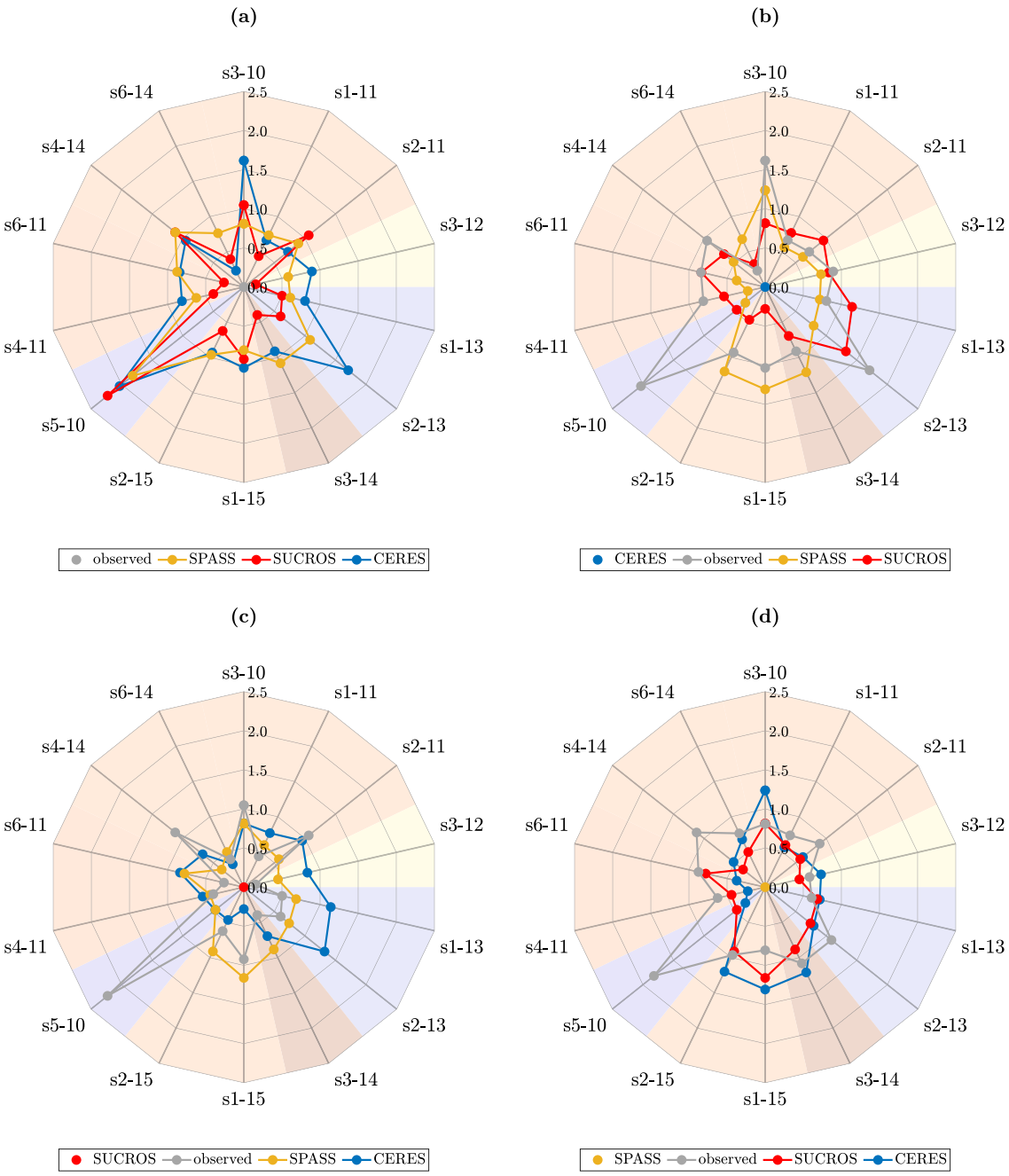


Fig. 4 Radar charts showing the energy distance between all models and observations based on yield predictions. In each subplot, one of the models or the observations are centered. Each colored line represents the distance of one model or the observations to the data set in the center. Each axis represents one site-year (abbreviated as, e.g. "s1–11" for site 1, year 2011). Segment colors resemble the annual weather conditions: hot and dry (red), average (yellow) to cold and wet (blue)

As can be seen from Figs. 5 and 11, different clusters are formed for different site-years. This shows that both, model similarity and goodness-of-fit, vary depending on the site-years.

Visualizing model similarities using heatmaps

In the heatmaps shown in Figs. 6 and 12, small values are represented by light colors and large values by dark colors. To dissect the individual components that the energy distance consists of, we plot its constituent parts $\mathbb{E}||X - Y||_2$ and $\mathbb{E}||X - X'||_2$ separately. The main diagonal entries represent the spread within the predictive distribution of a single model $\mathbb{E}||X - X'||_2$, while the off-diagonal entries represent the similarity of two objects $\mathbb{E}||X - Y||_2$.

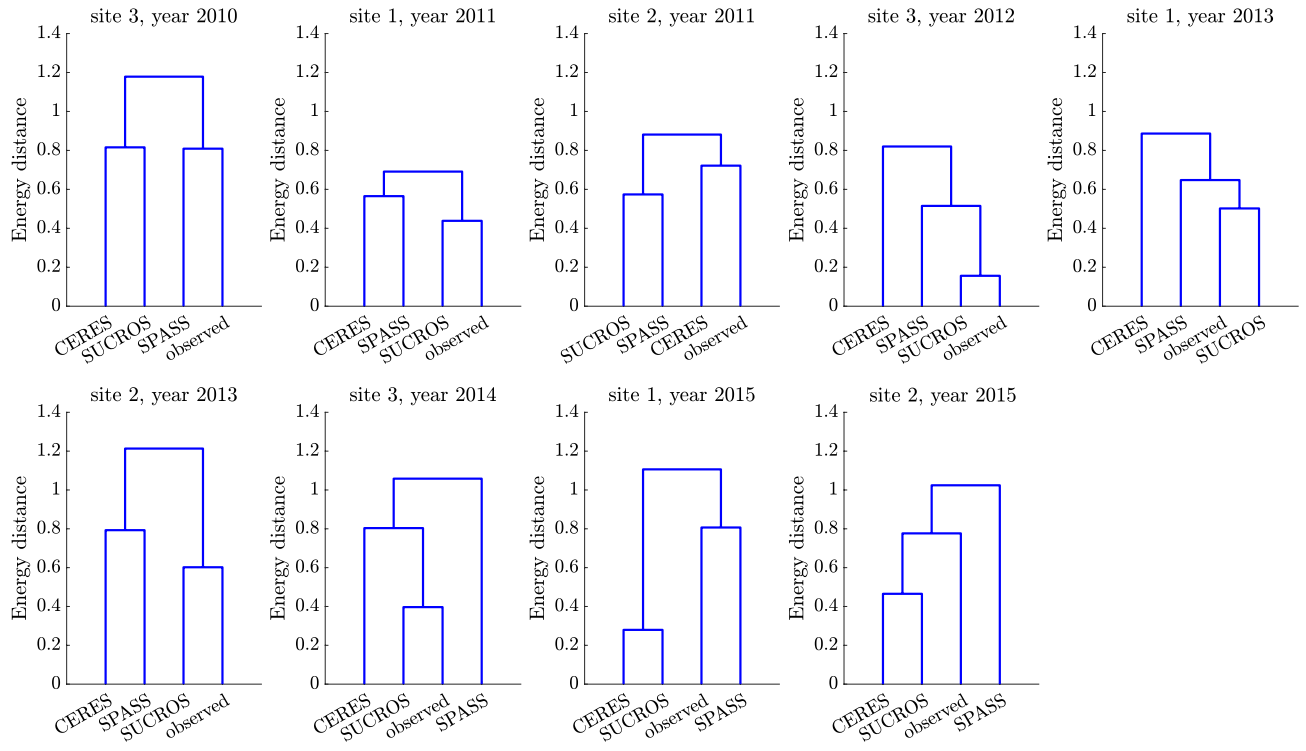


Fig. 5 Dendrograms showing the energy distance between models and observations based on yield predictions, sites 1-3

From the color-coding, we can intuitively see that the highest dissimilarity among the models occurs for site 3 in 2010 and 2014, as well as for sites 1 and 2 in 2015, as these heatmaps are overall darker than the others. In the same manner, it is immediately obvious that SPASS clearly differs from the other models and the measurements. This is neither visible in Fig. 4 (radar charts), nor in Fig. 5 (dendrograms). The last columns/rows show the goodness-of-fit to the observations. Here, it is clearly visible that SUCROS performs best and SPASS performs worst for most site-years.

Comparison of the visualization methods

Table 2 summarizes the properties of the three visualization methods and their applicability for different use cases.

With heatmaps and dendrograms, all objects (here: models) can be compared at one glance, however, only for one condition (here: site-year). Therefore, for comparing many models under specific conditions (e.g. per site-year), heat maps and dendrograms are suitable. In contrast, radar charts are useful for comparing one object (e.g. measurement data) to a small number of other objects (e.g. models) under many different conditions (e.g. site-years).

Dendrograms make it easy to identify clusters. Such clusters can, for example, indicate settings in which all models are similar, but are far from measurements. Such a case may point to the fact that an important process was not considered in any of the models. Examples for such a setting are the yield predictions for site 5, year 2010 and site 4, year 2014 (see Appendix B1). Of course, we can also see from the density functions in Fig. 3 or the radar chart centered on the observations (Fig. 4) that none of the models fits the measurements for this site-year well. However, the analysis based on density functions is only trivial in 1D cases such as the exemplary yield predictions used here. Imagine if we wanted to compare higher dimensional predictions. In such cases, this task would be much easier using dendrograms based on probability metrics. Also the analysis based on radar charts would be less convenient for identifying clusters: we can tell from the radar chart centered on the observations that all models are far off. However, we cannot tell based on the same radar chart whether the models are close to each other. To get this insight, we would have to analyze several radar charts with different models being centered.

Summary regarding the Expert-N model set

Summarizing the comparison of the Expert-N model set, SUCROS performs best in predicting yield based on the similarity of its prior predictive distribution and the distribution of the observations. The highest similarity among the models is between SPASS and SUCROS, while the biggest differences appear for SPASS and CERES. Our analysis confirms that yield predictions vary significantly between different models as Asseng et al. (2013) and Palosuo et al. (2011) found in earlier studies.

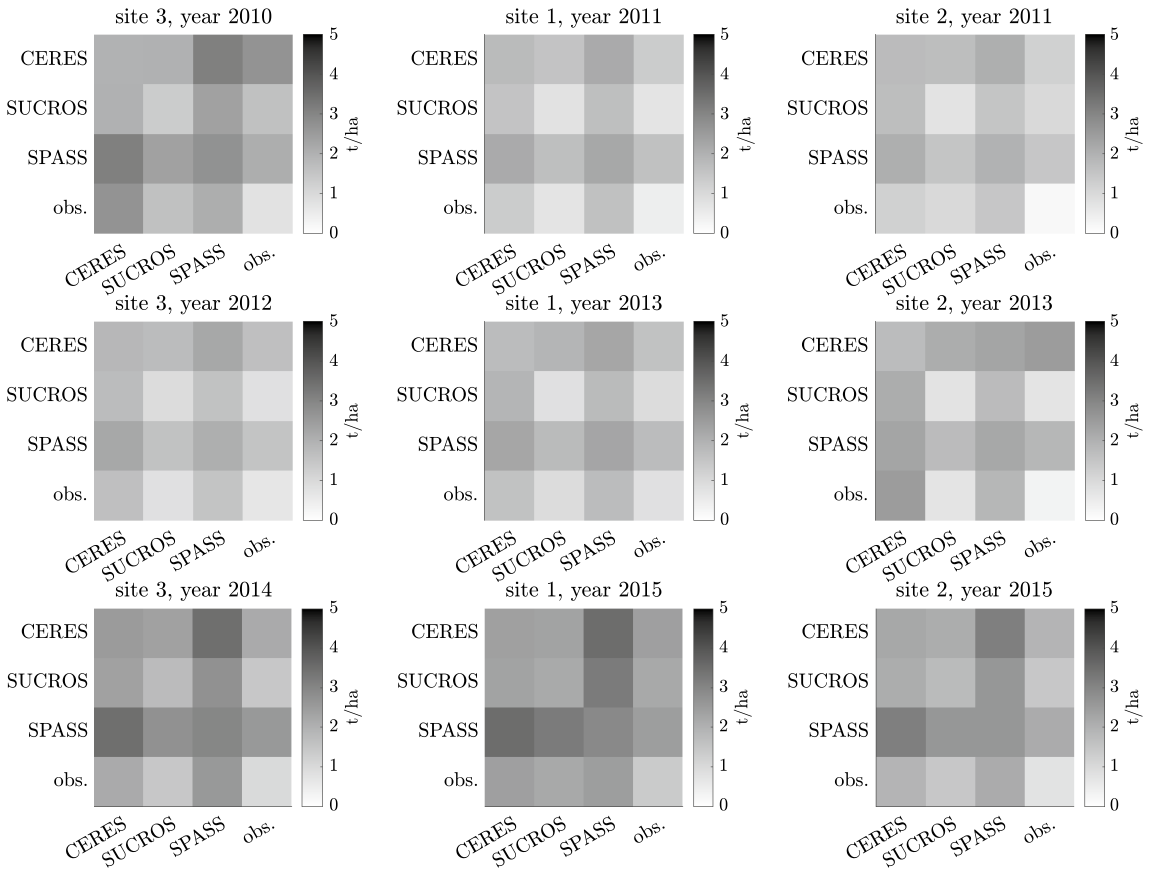


Fig. 6 Heatmaps reflecting the similarities between models and observations based on yield predictions, for sites 1–3. The color-coding represents the values of the individual components of the energy distance: $\mathbb{E}||X - Y||_2$ (off-diagonal entries) and $\mathbb{E}||X - X'||_2$ (main diagonal entries)

Table 2 Comparison of the visualization methods

Visualization method	Radar charts	Heat maps	Dendrograms
Type	Axis-based	Matrix-based	Hierarchy-based
Comparison of many objects	×	✓	✓
Comparison of many conditions	✓	(✓)	×
Easy identification of clusters	×	(✓)	✓
Color-coding possible	✓	(✓)	(✓)
Variation within and between objects	×	✓	×

The checkmark means that the method is well suited, the checkmark in parentheses means that the method can be used for the task in certain cases, but the visualization might become overloaded

Using the energy distance to aggregate the information and visualizing it with radar charts, dendrograms, and heat maps have been found insightful when inspecting influencing factors such as different sites or different weather conditions.

Analysis of the in-season variable phenology

After analyzing the end-of-season variable yield, we now focus on time series of the in-season variables phenological development stage (BBCH) and leaf area index (LAI, "Analysis of the in-season variable leaf area index"). Exemplary plots of the BBCH time series for site 1, year 2011 are shown in Fig. 7. The corresponding plots for all site-years are provided in the Appendix (Fig. 25).

We start with a qualitative analysis of the time series in "Predictive distributions". Next, we quantify how similar or different the models behave at each daily time step by calculating the energy distance between the predictive distributions in "Energy distance-based similarity analysis". We compare this between-model distance to the spread within

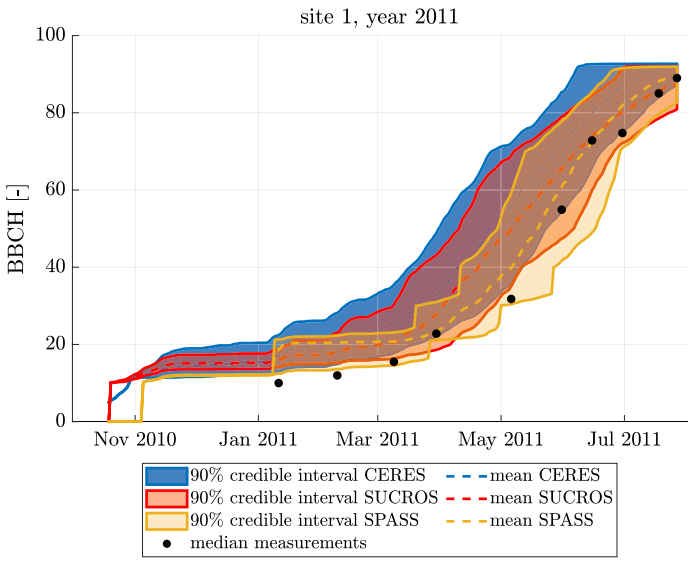


Fig. 7 Time series of phenology predictions for site 1, year 2011. The shaded intervals represent the 90% credible intervals. The points represent the median of the replicate measurements

each model. This spread is calculated as the square root of the mean Euclidean distance between all samples of the predictive distribution (see Eq. 1).

In contrast to the analysis of the yield predictions, we do not assume a distribution for the measurement errors for BBCH or LAI. Therefore, we use the median of the replicates and calculate the energy score, i.e. the counterpart of the energy distance for comparing distributions to a single observation.

Predictive distributions

From Fig. 7 we can see that the prior predictive distributions for the development stage generated by SPASS show a very small variance until January, whereas the predictions of CERES and SUCROS initially have little variance, but the spread increases already in November. This can be observed for all site-years (see Fig. 25).

Comparing the predictions to the measurements shows that early BBCH stages are usually overestimated by all three models. Starting approximately in March, when the development stage reaches values of 20, the predictions become more accurate.

The mean predictions of SPASS are closest to the measurements for most site-years, while particularly CERES, and to a lesser extent SUCROS, overestimate earlier phenological development. The discernible steps in the SPASS simulations are a direct result of and consistent with the model structure: in contrast to the two other models, which represent the secondary growth stages as fractions of the temperature sums required for each principal growth stage (BBCH = 10–20, 20–30, 30–40, …), SPASS simulates the secondary growth stages during the early development (BBCH = 10–40) based on the number of emerged main stem leaves (BBCH = 11, 12, …), tillers (BBCH = 21, 22, …), and main stem nodes (BBCH = 31, 32, …) Wang and Engel (1998). Therefore, in this model simulated BBCH may not be a continuous function of time. For example, if only five main stem leaves have unfolded by the day on which the principal growth stage “tillering” (BBCH = 20) has been reached, a discontinuity from BBCH = 15 to BBCH = 20 would be simulated. The model behaves similarly with respect to the number of tillers on day of principal growth stage “stem elongation” (BBCH = 30) and the number of nodes at principal growth stage “booting” (BBCH = 40).

We note that the better predictions of BBCH by SPASS are in contrast to the worst performance in yield. We can put these results into perspective with the yield predictions. While our belief about plausible parameter ranges of SUCROS and CERES led to an early onset of BBCH development, on average, the grain filling period from anthesis (BBCH = 60) until maturity (BBCH = 90) is longest. The price of this is a worse match in BBCH, in contrast to SPASS. While SPASS reaches maturity approximately similarly as SUCROS, it is the shorter grain filling duration in SPASS that results in a tendency for lower yields, shown by the heavy tails on the left of the predictive distribution functions (see Figs. 2, 3).

For the stated reasons, this delayed BBCH development (i.e., not-achieved maturity, leading to lower grain yields) does not occur in CERES. From this insight, we can update our formulation of plausible parameter ranges by ensuring that the grain filling rate in the SPASS model parameters is increased. Similar updates are possible for SUCROS and CERES, where a delayed BBCH would have been matched with an increased grain filling rate. Since, in practice, fertilization dates are co-informed by BBCH, this result is of great significance to enhance the prior predictive capabilities of the models for these types of environments.

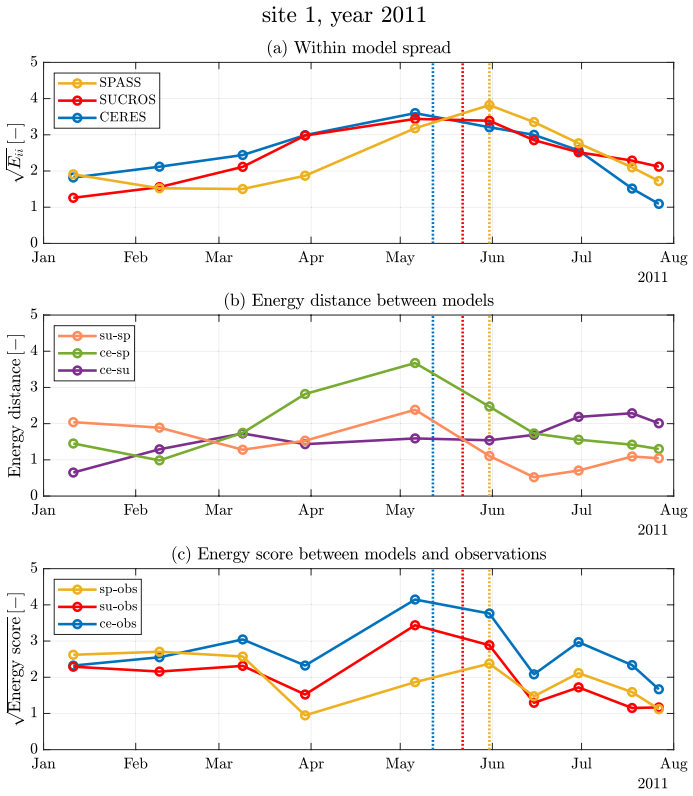


Fig. 8 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on phenology predictions, site 1, year 2011. The dashed lines indicate the date when the mean predictions reach BBCH = 60

Energy distance-based similarity analysis

Figure 8 (a) shows time series of the spread of each model, (b) the energy distance between pairs of models and (c) the energy score between models and observations. The corresponding figures for all other site-years are provided in the Appendix (Figs. 19, 20, 21, 22, 23 and 24).

From Fig. 8 we can see phenomena that can be observed in most site-years: Within-model spreads of phenology predictions (Fig. 8a) show an increasing trend until June and decrease steadily thereafter. This is to be expected since all models aim at reaching fill plant maturity by the time of harvest. While the curves for CERES (blue) and SUCROS (red) are similar, the one representing SPASS (yellow) is often shifted towards a later maximum in July and therefore shows the highest spread in its predictions at harvest date. From this, we can identify that the largest predictive uncertainty occurs around BBCH = 60, although it is one of the most important predicted stages. Within all models, the anthesis date at BBCH=60 is very important, as it marks the point at which grain filling starts. In principle, a very late start and short grain filling period (visible as steep slopes in the curves after BBCH=60 in Fig. 7) can be compensated with implausibly high grain filling rate parameters, such that reasonable yields can nevertheless be simulated. This can be achieved in all of the models.

In Fig. 8b, we can analyze the distance, i.e. dissimilarity between pairs of models. The distance between CERES and SPASS (green curve) is the highest during most phases for all site-years. After reaching the maximum distance in May, it decreases again. The distance between SUCROS and SPASS (orange curve) shows a similar development, however, the distance between these models is smaller most of the time. The distance between CERES and SUCROS (purple curve) does not show this characteristic maximum in May, rather it increases more or less steadily over the season and thus, the difference between CERES and SUCROS becomes the highest at the harvest date.

The curves in Fig. 8c can be interpreted as the models' goodness-of-fit to measurements. A low energy score means a low distance to the observations and hence a good model performance. The time series of CERES and SUCROS show similar behavior with a maximum energy score in May to June, followed by a decline. For most site-years, CERES has the highest energy score. Although in some cases SPASS starts with the highest energy score, for most site-years, it is closest to the observations during most phases. Hence, from a model selection perspective, SPASS would be considered best overall. Nonetheless, the analysis with the energy score clearly shows that, especially in certain time windows, employing additional alternative models increases reliable predictive coverage.

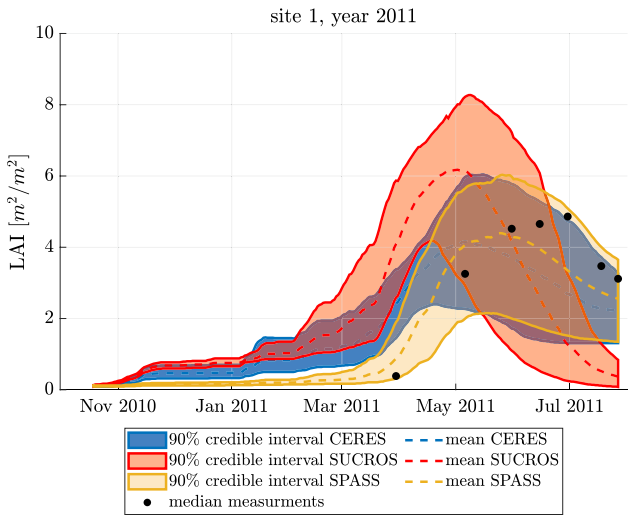


Fig. 9 Time series of LAI predictions for site 1, year 2011. The shaded intervals represent the 90% credible intervals

Analysis of the in-season variable leaf area index

Predictive distributions

After the analysis of BBCH, we now study the in-season variable LAI. Exemplary plots of the time series for site 1, year 2011 are shown in Fig. 9. The corresponding plots for all site-years are provided in the Appendix (Fig. 26).

The adopted measurement technique (Weber et al. 2021) for field observations of LAI does not differentiate between green leaf (i.e. photosynthetically active) and dead leaves. After the onset of leaf senescence, which dominates the LAI evolution after maximum LAI, we can consider that the measured LAI values contain an unknown amount of green leaf as well as dead leaves. However, the modeled LAI is green leaf LAI. For LAI, we can see clear differences between the predictions of the three models: CERES and SUCROS overestimate the LAI in the initial phase, however, the predictions of CERES do not drop as significantly as the ones of SUCROS. As for the phenological development, SPASS can describe the measured data most accurately.

In all cases, the peak of the median simulated LAI is much earlier than the peak in the measurements, sometimes by far. Since green leaf LAI of winter wheat is 0 at harvest, we see that, out of the three models, only a few individual simulation runs of SUCROS achieve this. Recall that the SUCROS wheat model does not simulate the vernalization of winter wheat. In other words, it is a summer wheat model, providing a feasible explanation of the, comparatively, very early development of LAI. This overestimation leads to a premature maximum and underestimates the measurements in the decreasing phase during senescence.

Figure 9 indicates that the prior predictive of SUCROS at harvest is close to zero, however with both an early and large peak in simulated LAI. Both CERES and SPASS are closer to the data but do not reach a green leaf of 0 at harvest. For SPASS this could relate to the fact that the simulated phenology had not fully reached maturity by the time of the harvest. In other words, if in the model the harvest date had been set to full maturity, and not to the farmer-reported harvest date, we would surely observe a further decrease in LAI. While the development stage of SPASS is "too slow", we can learn that the senescence of CERES is not fast enough.

Energy distance-based similarity analysis

Similar to BBCH, the within-model spread of the LAI predictions in Fig. 10a increases slowly until June, followed by a decrease until harvest. The curves representing CERES and SPASS are relatively similar. SUCROS shows the largest spread for most of the simulation time until it declines starting in June. At the harvest date, it is mostly SPASS that has the highest spread in its predictions.

In Fig. 10b, the energy distance between the models shows that, during most of the seasons, the predictions of CERES and SPASS are the most similar ones. The time series of the energy distance between SUCROS and CERES and between SUCROS and SPASS are similar, with the ones of SUCROS and CERES being usually lower (i.e. the models are more similar) and showing a minimum in June–July, followed by a rising phase until harvest.

Figure 10c shows that SPASS has the lowest distance to the observations for most months, which means it performs best. The curve representing CERES' energy score ranges in the middle, and SUCROS performs worst, having the highest distance. The case of SUCROS highlights one of the major benefits of analyzing model predictive distributions using energy distance and energy score: While SUCROS shows the worst predictions according to Fig. 10c, it provides these with the highest confidence in the time from July to August. Energy statistics support this insight in a straightforward way on an easily-interpretable scale.

Compared to the energy-statistics-based analysis of BBCH (Fig. 8), the within-model spreads of the LAI predictions (Fig. 10) are smaller and increase less during the growing season. From the energy distance between the models, we can observe that CERES and SPASS are the most dissimilar models with regard to BBCH predictions, while they are the most similar ones considering LAI predictions. Here, the biggest differences are between the LAI predictions of SUCROS and SPASS. The smallest energy score and, hence, the best goodness-of-fit during most phases was calculated, for both quantities of interest, by SPASS. The worst performing model regarding LAI predictions is SUCROS, while considering BBCH predictions, it is CERES.

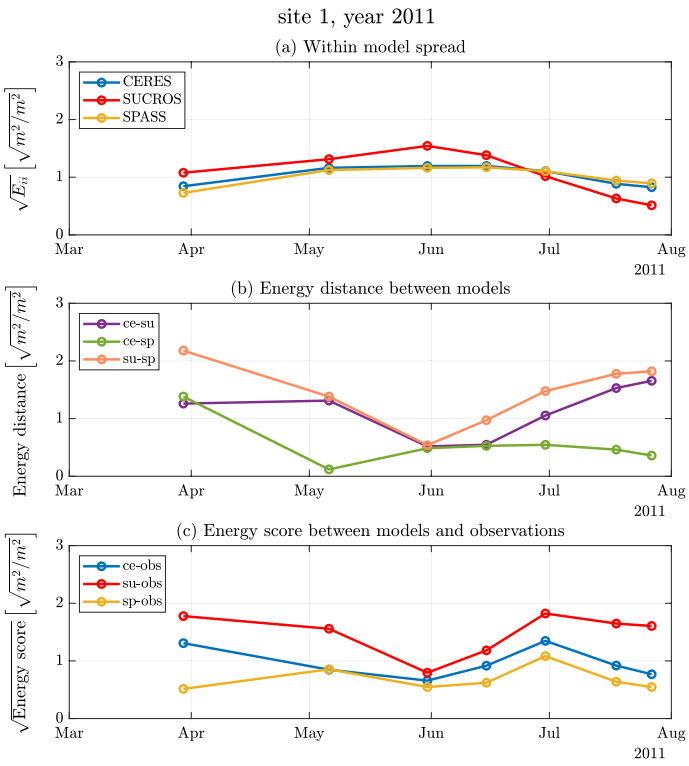


Fig. 10 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on LAI, site 1, year 2011

Our analysis of the prior predictive distributions revealed that a model that performs well during season might still end with an imprecise yield prediction: During the season, SPASS is best in predicting LAI and BBCH, whereas CERES and SUCROS clearly deviate from the observations, especially in the case of LAI. In predicting yield, however, SUCROS performs best, while SPASS is worst due to its very broad prior predictive distribution that covers even very low values with a relatively high probability. This confirms what was observed in the study by Martre et al. (2015). They compared the goodness-of-fit for calibrated models: a model that cannot reproduce in-season measurements well might do a better job in predicting end-of-season variables. Given our analysis, one could add to this statement that also the reverse can be true, i.e. a model that performs well during the season might still fail to predict yield reasonably.

In addition, the similarity among the models was not consistent across different variables: while the LAI predictions of CERES and SPASS are the most similar ones and the ones of SUCROS and SPASS are the most dissimilar ones, the opposite is true considering the yield predictions.

Summary and conclusions

We analyzed the similarity of predictions by the three plant growth models CERES, SPASS and SUCROS and their goodness-of-fit to observed data in a probabilistic framework. The goal of this study was to find methods for gaining deeper insights into the model set. An intuitive understanding of similarities between the models and the measurements can help model developers to improve both the individual models and the multi-model methods. The presented method can be used to identify different model settings, e.g. situations in which all models form a cluster while being distant from the measured data. This may indicate that all models are highly similar and that a relevant process is not considered in any of the models. An intuitive visualization of model similarities can guide the multi-model process, e.g. when it comes to assigning model weights for averaging. Therefore, we propose to combine specific visualization methods that make modelers aware of the (dis)similarities in the predictions of the considered model set. Each method highlights another piece of information and adds to a comprehensive overview of the considered model set.

The analysis is based on so-called energy statistics introduced by Rizzo and Székely (2016). The energy distance between the probabilistic predictions is used to quantify model similarities. With the same method, we can also assess model performance by calculating the energy distance between model predictions and noisy measurements. For comparing probabilistic model predictions to deterministic observations, the so-called energy score is used. It acts on the same scale as the energy distance making both intuitively comparable. Therefore, energy statistics proved to be widely applicable, as energy distance and energy score can be used

jointly to compare two probability distributions as well as a probability distribution and a deterministic reference.

Our results confirmed that "there is no single best model" (Hagedorn et al. 2005; Palosuo et al. 2011; Martre et al. 2015): none of the investigated models performed consistently better or worse than the others when considering different variables. While SPASS showed the best goodness-of-fit regarding in-season variables LAI and BBCH, its overly wide yield predictions lead to poor performance for this end-of-season variable. Therefore, combining the models in an ensemble might indeed give more robust predictions as a broader range of possible predictions is covered.

Generally, we suggest analyzing model similarities when using multi-model ensembles, as redundancies in the model set lead to an overly high weight of certain predictions and therefore, model weights should be diluted (George 2010; Garthwaite and Mubwandarikwa 2010). Similar to the results regarding goodness-of-fit, we also found that model similarities vary for different variables: two models that gave similar predictions for one variable showed clear differences in predicting another one. Therefore, no general dilution priors can be defined for this model set. Rather, they need to be chosen depending on model similarities for each quantity of interest.

We also investigated whether model similarities or performance are dependent on the weather conditions during the growing season. To this end, we used radar charts to visualize the similarities and color-coded them according to the wetness or dryness of the respective site-year. Although there was no apparent effect of the weather conditions on the model predictions visible, we suggest this approach of visualization to be studied further. We assume that, for other scenarios and model sets, this might be a straightforward tool to display the influence of different boundary conditions on the prediction accuracy and similarity of models. By assessing the within-model spread, the distance between the models, and the goodness-of-fit on the same scale, we can gain a better understanding of the model set.

Our study was based on prior predictions, i.e. the models have not been calibrated. There are two main reasons for this: (1) (not only) in the crop modeling community, different groups use different calibration approaches (Wallach et al. 2020) and hence, there is a lack of consistency. (2) Model structural errors are often compensated by choosing non-physical parameters (e.g. Wallach 2011). This leads to good model performance for the variable the model has been calibrated on, but poor performance for others. Therefore, we support the suggestion of Vogel and Sankarasubramanian (2003) to validate the model structure prior to calibration. As in any Bayesian framework, a subjective choice of prior distributions based on expert knowledge is needed. Future research should assess the sensitivity of the analysis regarding the priors. Another promising way to go is the assessment of structural model similarity, e.g. based on information-theoretic methods as done by Bennett et al. (2019). As our analysis suggests that a combination of the individual models into an ensemble prediction might yield more robust results, our introduced model evaluation workflow might also inform different model combination methods and weighting schemes of future applications.

Appendix 1: Model description

See Tables 3, 4, 6, 5, 7.

Table 3 Characteristics of the three models, adapted from Asseng et al. (2013), supplementary material (Table S2)

	CERES	SPASS	SUCROS
Leaf area/light interception	Simple	Detailed	Detailed
Light utilization	Radiation use efficiency approach	Gross photosynthesis-respiration	Gross photosynthesis-respiration
Yield formation	Tot. (above-ground) biomass, number of grains	Number of grains, partitioning during reproductive stages	Partitioning during reproductive stages
Phenology	Temperature, photoperiod (day length), vernalization	Temperature, photoperiod (day length), vernalization	Temperature
Root distribution over depth	Exponential	Exponential	Exponential
Environmental constraints	Water limitation, *N* limitation	Water limitation, *N* limitation	Water limitation, *N* limitation
Type of water stress	Actual to potential evapotranspiration ratio, soil available water in root zone	Actual to potential evapotranspiration ratio, soil available water in root zone	Actual to potential evapotranspiration ratio, soil available water in root zone
Water dynamics	Richards approach	Richards approach	Richards approach
Evapotranspiration	Penman–Monteith	Penman–Monteith	Penman–Monteith
Soil CN-model	CN model, 3 organic matter pools, microbial biomass pool	CN model, 3 organic matter pools, microbial biomass pool	CN model, 3 organic matter pools, microbial biomass pool
No. of cultivar parameters	7	5	2

Table 4 Description of model parameters varied and parameter bounds of the uniform prior where a = lower bound and b = upper bound, subscript 1 denotes the top-soil and 2 the sub-soil as defined in Table 1

Module	Parameter	Unit	a	b	Description
Soil water	α_1, α_2	cm^{-1}	0.002	0.03	van Genuchten α
	n_1, n_2	–	1.2	2.2	van Genuchten n
	$K_{s,1}, K_{s,2}$	cm d^{-1}	10	500	saturated hydraulic conductivity
	τ_1, τ_1	–	− 1	8	Tortuousity parameter
ET_a	kc_{ini}	–	0.2	1	Crop coefficient for the initial vegetation period (dev stage 0)
	kc_{mid}	–	0.5	1.8	Crop coefficient for the mid vegetation period (dev stage 0.75–1.5)
	kc_{end}	–	0.2	1	Crop coefficient for the end of the vegetation period (dev stage 2.0)
Nitrogen	Nit_1	d^{-1}	0.1	1	Nitrification rate
	Nit_2	d^{-1}	0.05	0.6	
	DeNit_1	d^{-1}	0.1	1	Denitrification rate
	DeNit_2	d^{-1}	0.0	0.01	
	MiLit_1	d^{-1}	0.01	0.1	Mineralization rate constant of the litter pool
	MiLit_2	d^{-1}	0.01	0.1	
	MiHum_1	d^{-1}	$1e^{-5}$	$1e^{-4}$	Mineralization rate constant of the humus pool
	MiHum_2	d^{-1}	$1e^{-6}$	$1e^{-5}$	
	MiMa	d^{-1}	0.01	0.1	Mineralization rate constant of the manure pool
Initial condition	$\theta_{ini,1}$	vol%	10	40	Initial soil profile water content
	$\theta_{ini,2}$	vol%	15	45	

Table 5 Description of SPASS model parameters and parameter bounds of the prior

Parameter	Unit	*a*	*b*	Description
PMAX	$kg_{CO_2}\ ha_{leaf}^{-1}\ h^{-1}$	38	45	Gross photosynthesis rate at light saturation and CO_2 340 ppm
LUE	$g\ J^{-1}$	0.55	0.7	Light use efficiency
TMINPS	°C	0	5	Minimum temperature for photosynthesis
TOPTPS	°C	20	25	Optimum temperature for photosynthesis
TMAXPS	°C	32	40	Maximum temperature for photosynthesis
PDD1	d	32	48	Duration from emergence to anthesis
PDD2	d	20	36	Duration from anthesis to emergence
VERN	d	24	46	Minimum value of vernalization days
PDL	–	0.01	0.25	Photoperiod sensitivity factor
DLOPT	h	18	20	Optimal photoperiod length
TMINDEV	°C	0	2	Minimum temperature of vegetative development
TOPTDEV	°C	22	26	Optimum temperature of vegetative development
TMAXDEV	°C	32	38	Maximum temperature of vegetative development
TMINDEV2	°C	2	6	Minimum temperature of reprod. development
TOPTDEV2	°C	26	30	Optimum temperature of reprod. development
TMAXDEV2	°C	38	45	Maximum temperature of reprod. development
G1	$\#g^{-1}$	24	35	Number of grains per stem weight at anthesis
SPCLW	$kg_{DW}\ ha_{leaf}^{-1}$	350	500	Specific leaf weight
RESR	–	0.36	0.45	Fraction of stem weight as reserves
G2	$mg\ grain^{-1}\ d^{-1}$	2.4	3.6	Maximum grain filling rate
NACCR	$mg_N\ grain^{-1}\ d^{-1}$	0.02	0.06	Nitrogen accumulation rate
REXT	$cm\ d^{-1}$	1.5	3.0	Maximum root extension rate
SPCRL	$cm\ g^{-1}$	8000	12000	Specific root length
RWUR	$cm^3\ cm^{-1}\ d^{-1}$	0.024	0.036	Maximum water uptake rate per root length
RNUR	$kg_N\ cm^{-1}\ d^{-1}$	0.006	0.01	Maximum nitrogen uptake rate per root length
RDMAX	cm	100	200	Maximum rooting depth
DVSSEN	–	1	1.3	Development stage at which senescence begins
RDRL	–	0.015	0.025	Relative death rate of leaves
RDRR	–	0.015	0.025	Relative death rate of roots

Table 6 Description of CERES model parameters and parameter bounds of the prior

Parameter	Unit	*a*	*b*	Description
P1D	–	0.001	0.008	Daylength coefficient
P1V	h	25	60	Inverse of optimum vernalization rate
PHINT	–	70	150	Phyllochrone interval
P1	°C d	170	400	Thermal Time from emergence-to terminal spikelet
P4	°C d	120	200	Thermal Time from end of pre-anthesis ear growth-begin of grain filling
P5	°C d	400	700	Thermal Time for grain filling (phase 5)
G1	$\#g^{-1}$	20	40	Number of grains per stem weight at anthesis
G2	$mggrain^{-1}\ d^{-1}$	1	4	Maximum grain filling rate
RWUR	$cm^3\ cm^{-1}\ d^{-1}$	0.01	0.1	Maximum water uptake rate per root length
RNUR	$kg\ ha^{-1}$	0.003	0.027	Maximum nitrogen uptake rate per root length

Table 7 Description of SUCROS model parameters and parameter bounds of the prior

Parameter	Unit	*a*	*b*	Description
PMAX	$kg_{CO_2}\ ha_{leaf}^{-1}\ h^{-1}$	38	45	Gross photosynthesis rate at light saturation and CO_2 340 ppm
LUE	$g\ J^{-1}$	0.55	0.7	Light use efficiency
TBASE1	°C	0	2	Base temperature for phen. dev. vegetative phase
TSUM1	°C d	700	1500	Temperature sum of vegetative growth phase
TBASE2	°C	2	5	Base temperature for phen. dev. generative phase
TSUM2	°C d	600	1400	Temperature sum of generative growth phase
LA0	$m^2\ plant^{-1} * 10{,}000$	0.45	0.7	Initial leaf area
RGRL	$°C^{-1}\ d^{-1}$	0.005	0.01	Relative growth rate of leaf area
G1	$\#g^{-1}$	24	35	Number of grains per stem weight at anthesis
SPCLW	$kg_{DW}\ ha_{leaf}^{-1}$	350	500	Specific leaf weight
REXT	$cm\ d^{-1}$	1.5	3.0	Maximum root extension rate
SPCRL	$m\ kg^{-1}$	8000	12,000	Specific root length
RDMAX	cm	100	200	Maximum rooting depth

Appendix B: Figures

See Figs. 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 and 26.

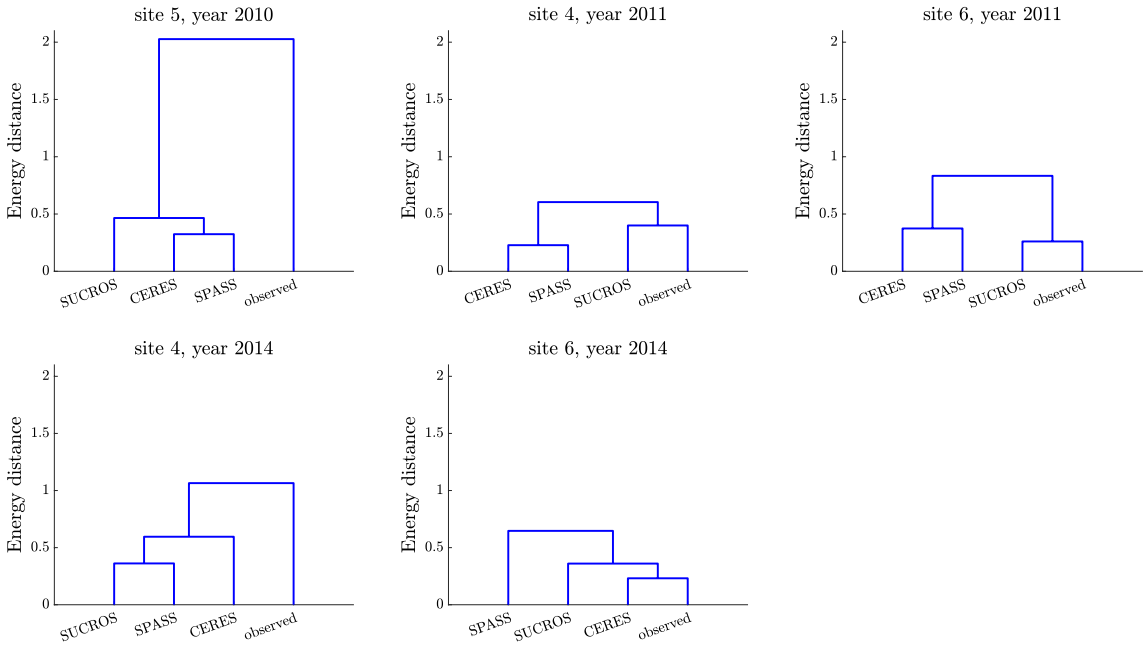


Fig. 11 Dendrograms based on the energy distance between models and observations (yield predictions)

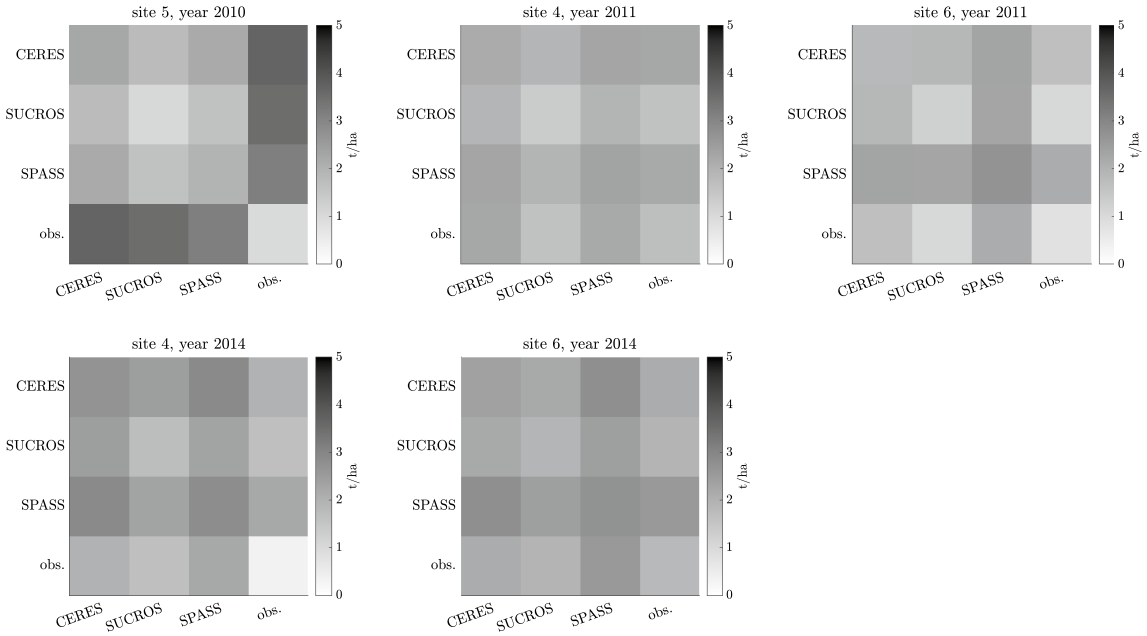


Fig. 12 Heatmap reflecting the similarities between models and observations based on yield predictions for site 4–6. The color-coding represents the values of the individual components of the energy distance: $\mathbb{E}||X - Y||_2$ (main diagonal entries) and $\mathbb{E}||X - X'||_2$ (off-diagonal entries)

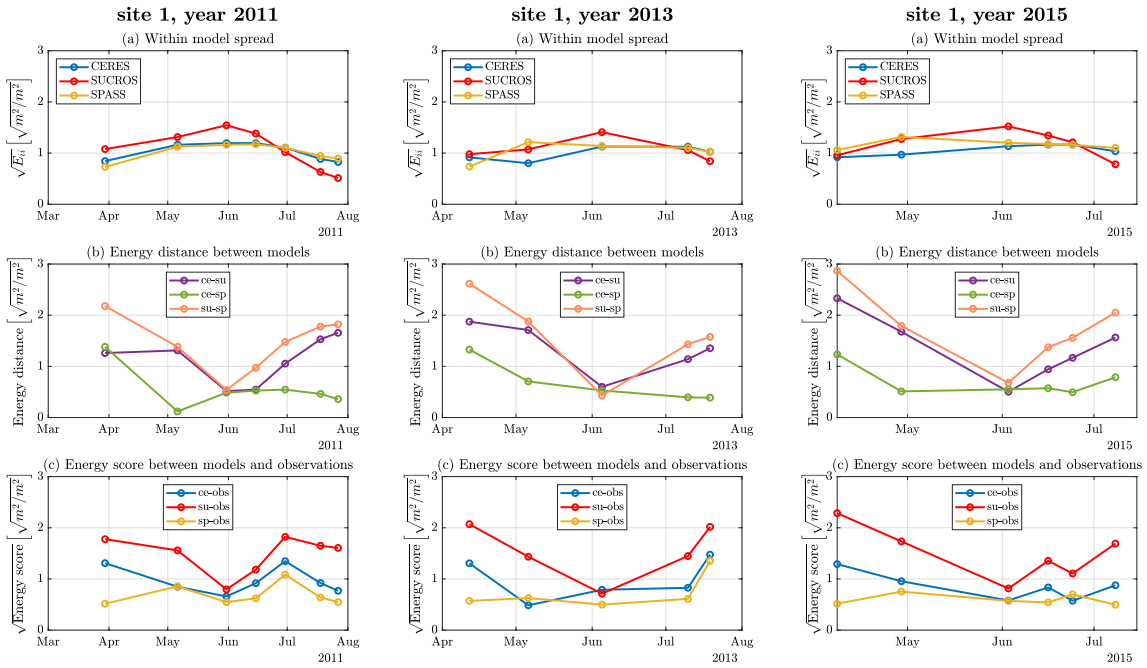


Fig. 13 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on LAI for site 1

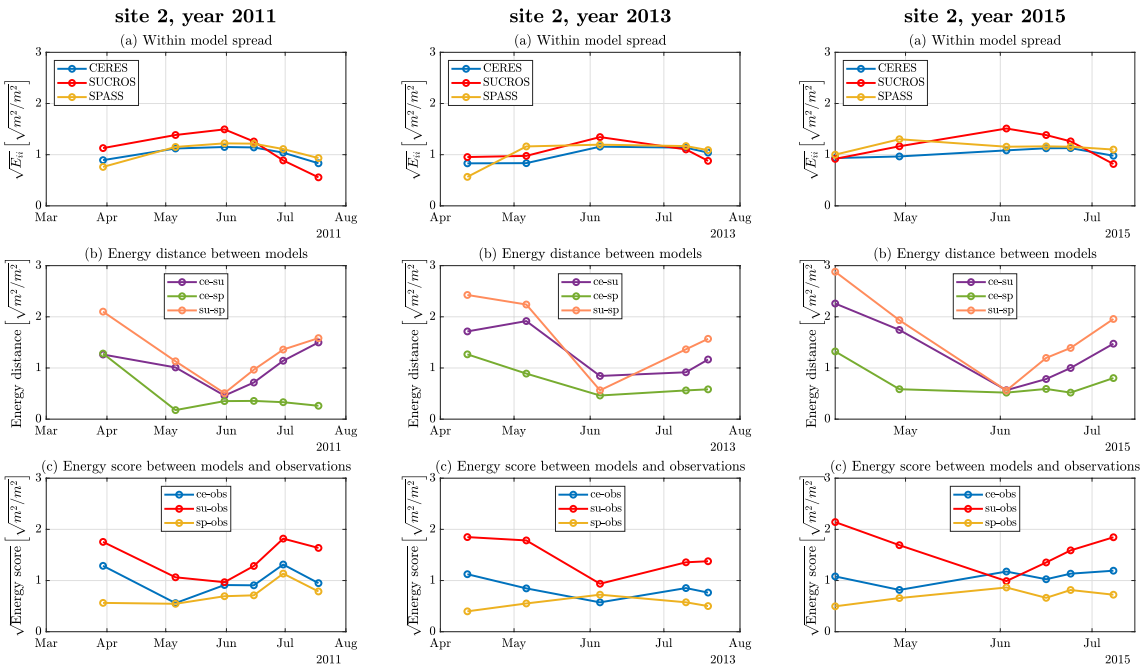


Fig. 14 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on LAI for site 2

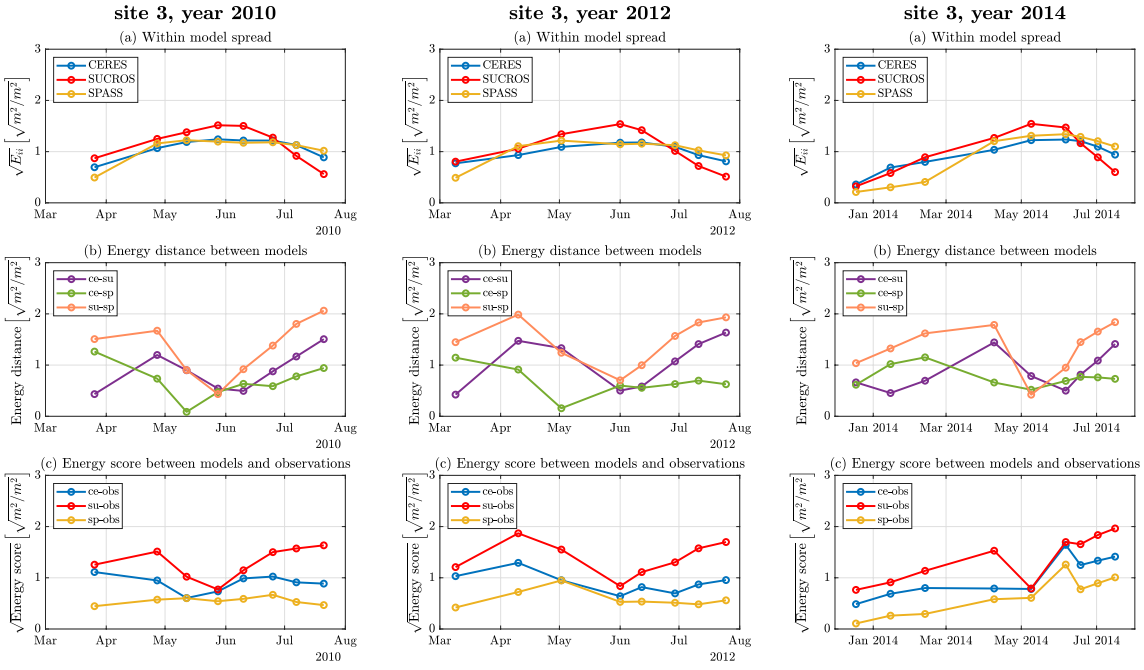


Fig. 15 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on LAI for site 3

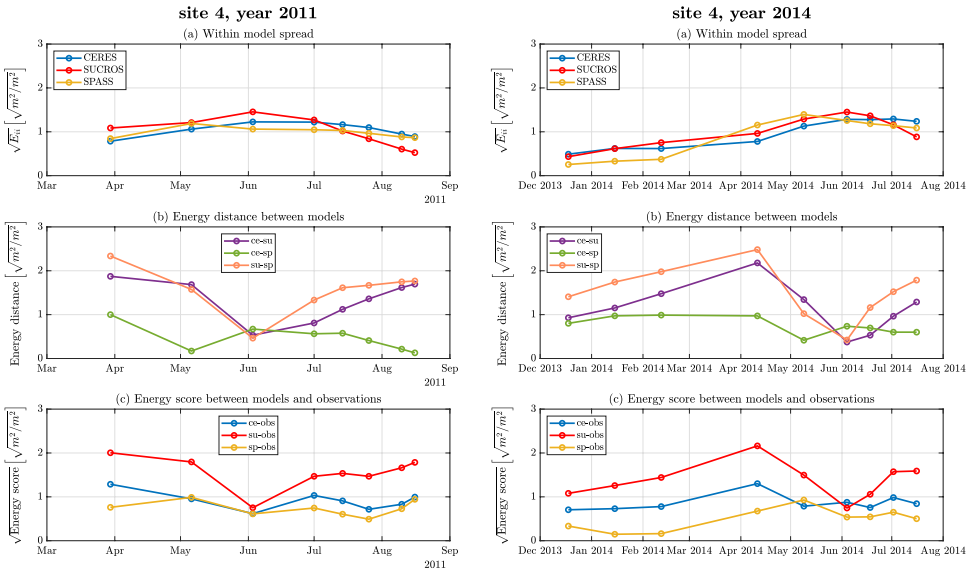


Fig. 16 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on LAI for site 4

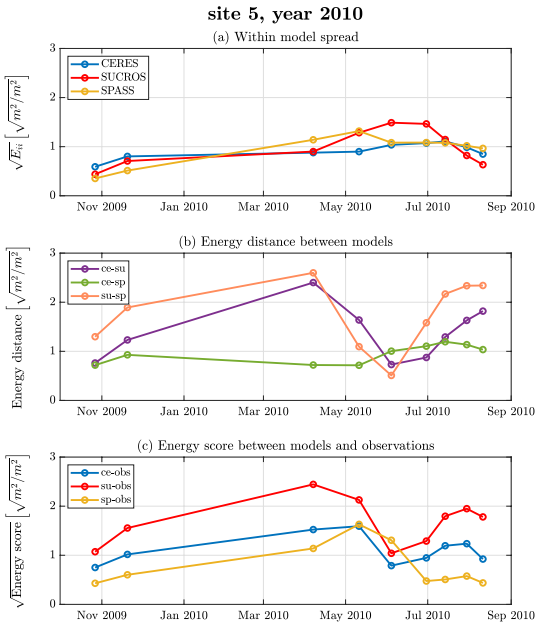


Fig. 17 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on LAI for site 5

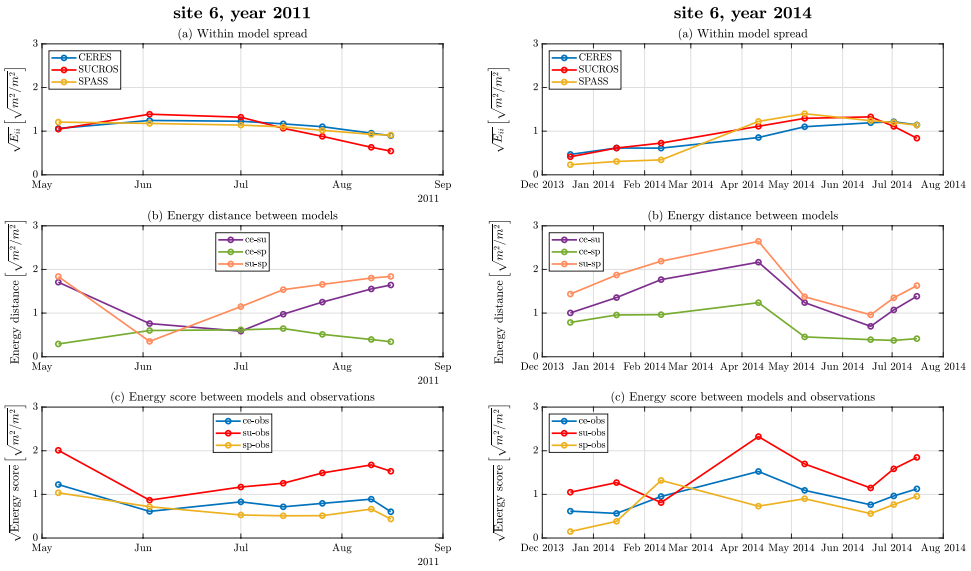


Fig. 18 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on LAI for site 6

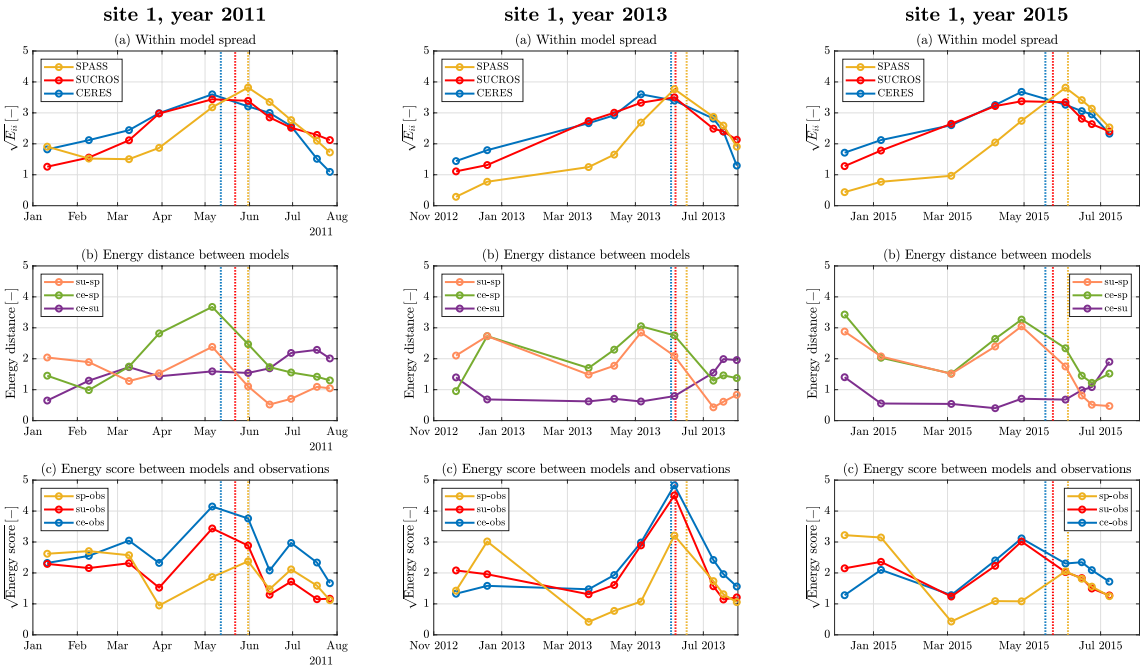


Fig. 19 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on phenology for site 1. The dashed lines indicate the date when the mean predictions reach BBCH = 60

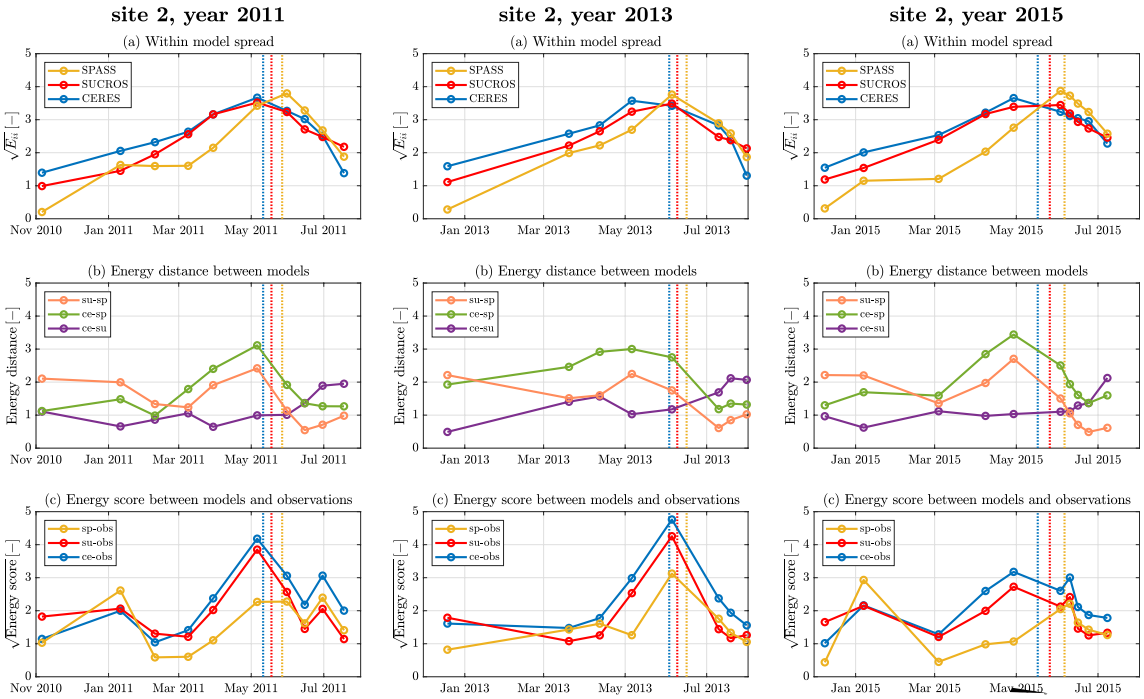


Fig. 20 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on phenology for site 2. The dashed lines indicate the date when the mean predictions reach BBCH = 60

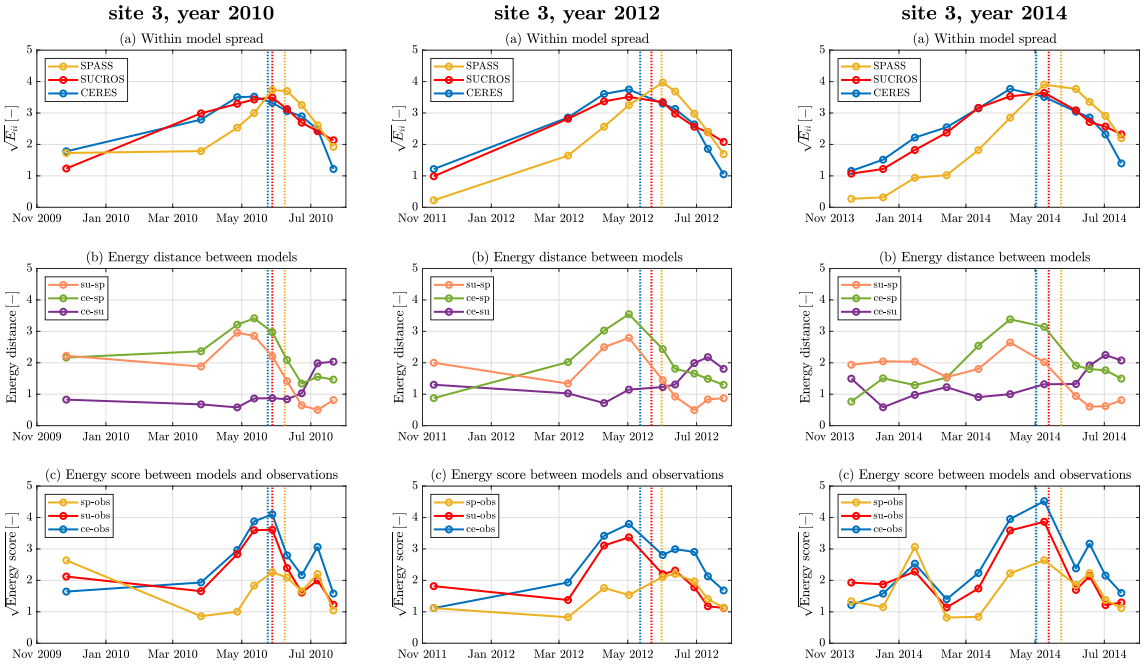


Fig. 21 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on phenology for site 3. The dashed lines indicate the date when the mean predictions reach BBCH = 60

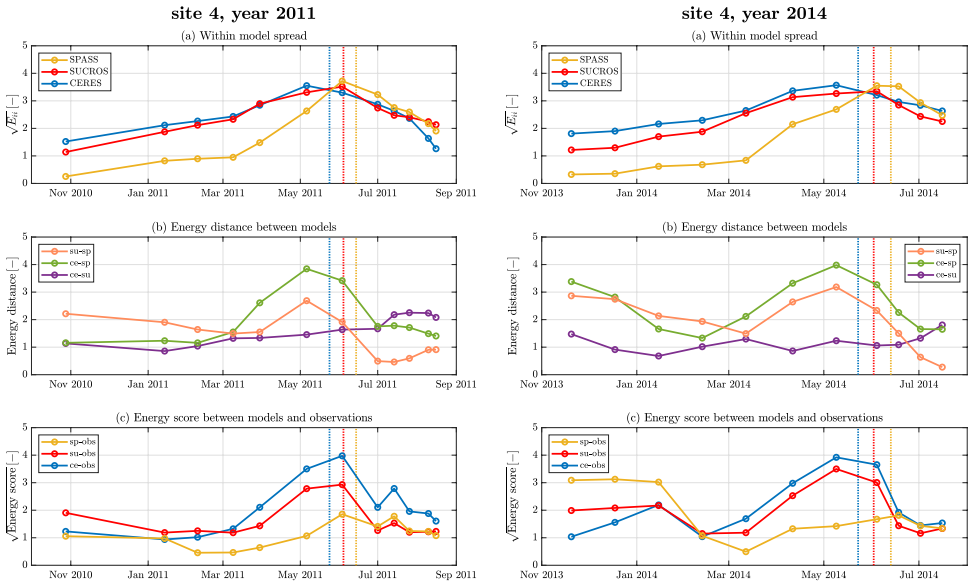


Fig. 22 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on phenology for site 4. The dashed lines indicate the date when the mean predictions reach BBCH = 60

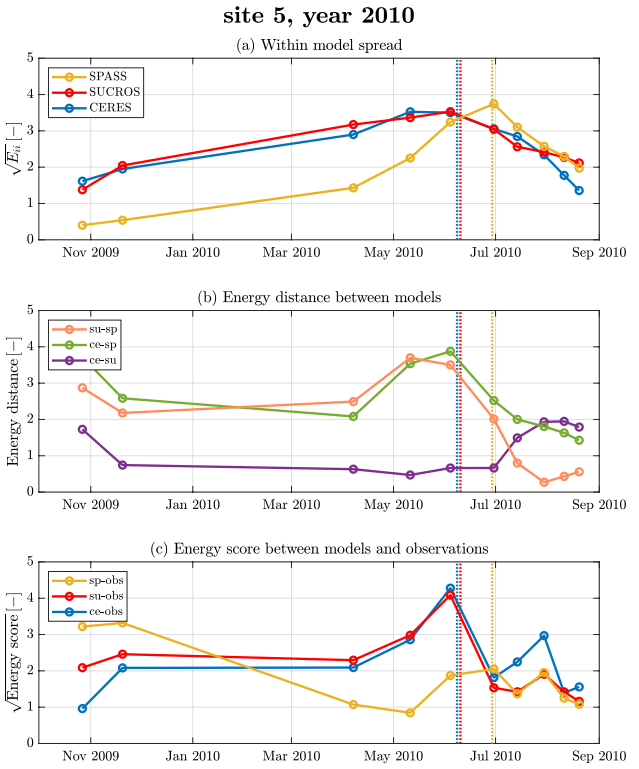
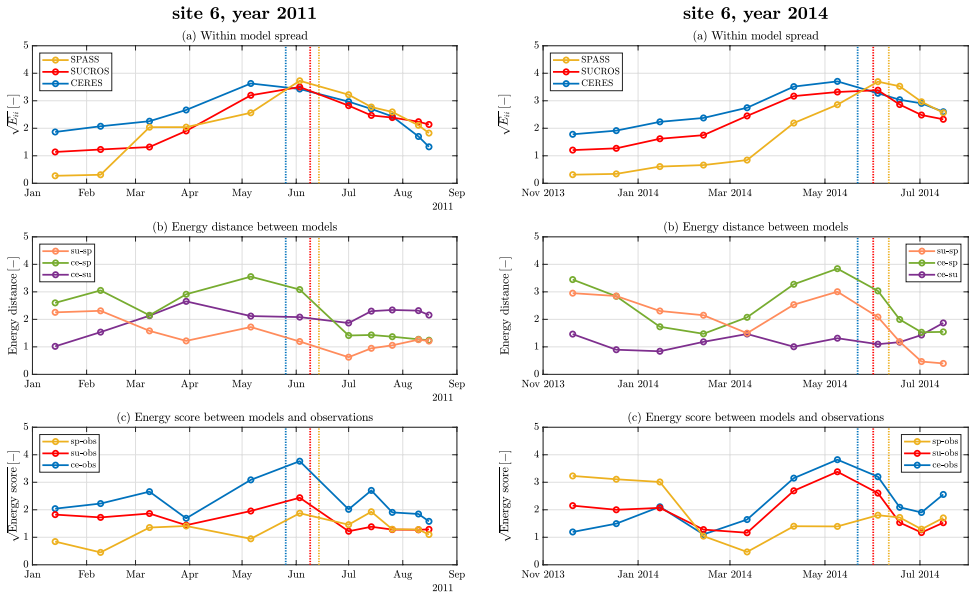


Fig. 23 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on phenology for site 5. The dashed lines indicate the date when the mean predictions reach BBCH = 60



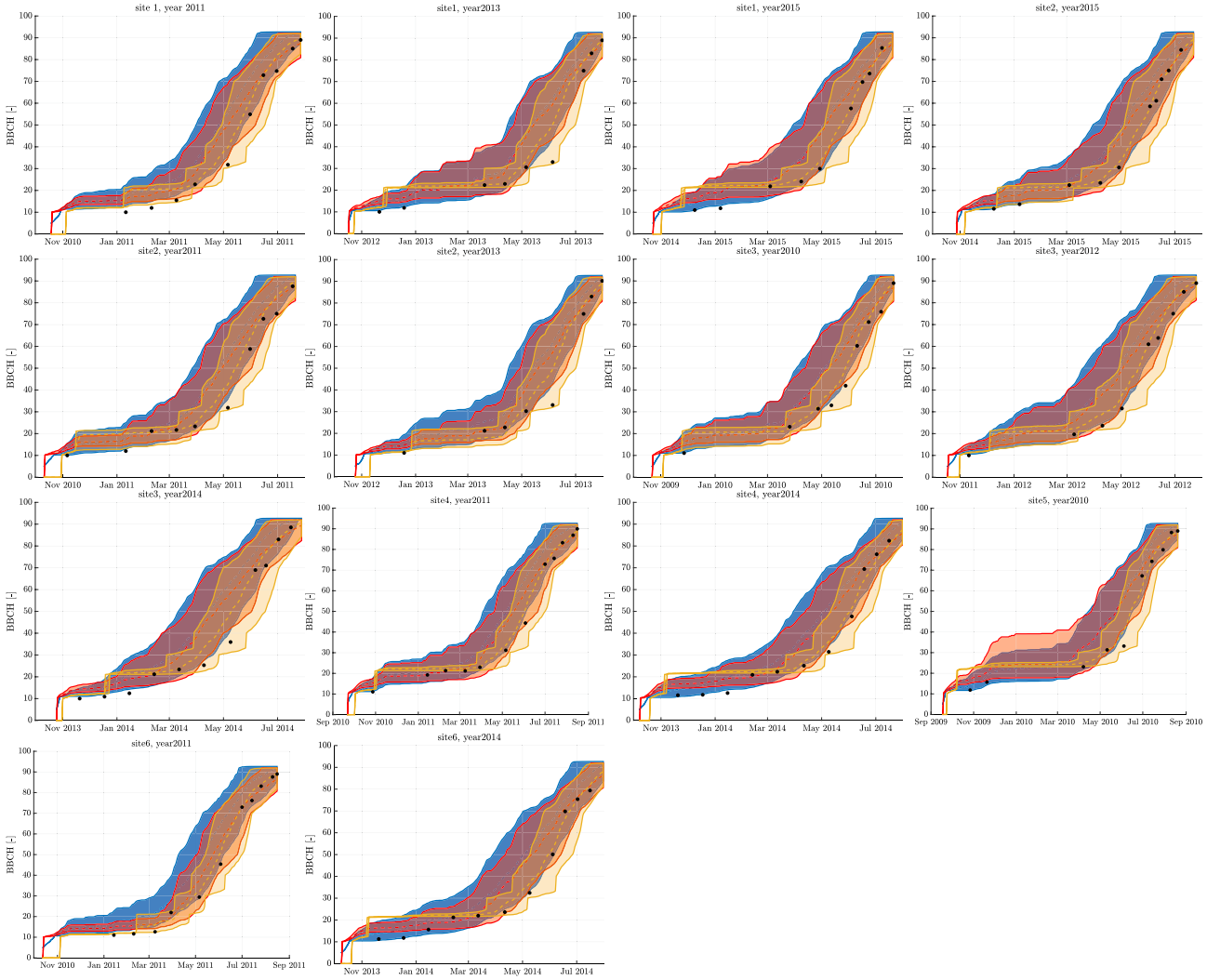


Fig. 24 **a** Within-model spread (square root of the mean Euclidean distance between the samples within each model) $\sqrt{E_{ii}}$, **b** energy distance d between pairs of models, and **c** root energy score $\sqrt{\text{ES}}$ between models and observations based on phenology for site 6. The dashed lines indicate the date when the mean predictions reach BBCH = 60

Fig. 25 Time series of phenology predictions for all site-years. The shaded intervals represent the 90% credible intervals (blue: CERES, red: SUCROS, yellow: SPASS), the dashed lines represent the model means and the black points represent the medians of the measurements

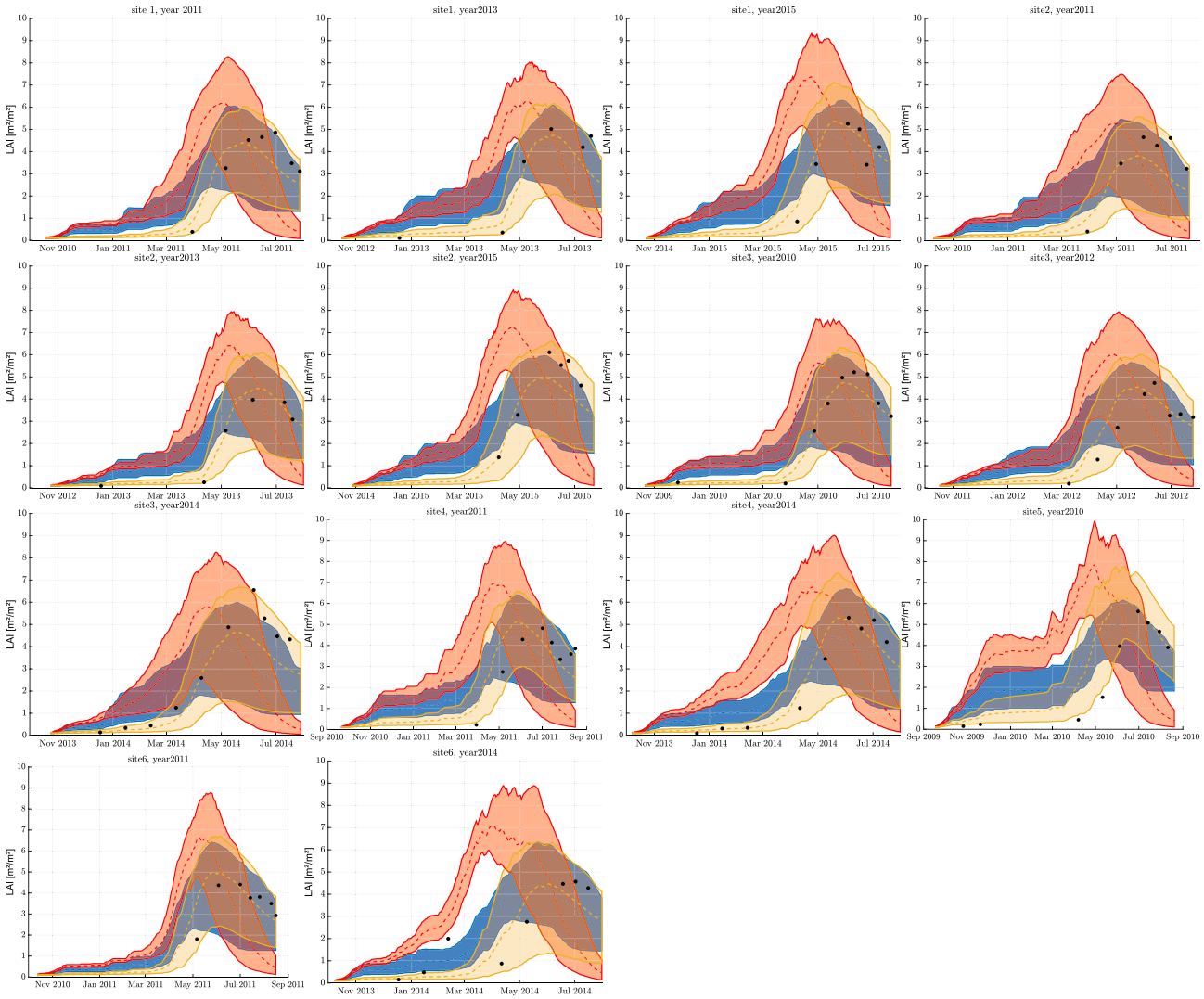


Fig. 26 Time series of LAI predictions for all site-years. The shaded intervals represent the 90% credible intervals (blue: CERES, red: SUCROS, yellow: SPASS), the dashed lines represent the model means and the black points represent the medians of the measurements

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) 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. We also thank Dr. Irene Witte for providing the Monte Carlo simulation results.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availibility Data and code can be downloaded here: https://bwsyncandshare.kit.edu/s/4gZZf4qEAA4TFMn.

References

Abramowitz G (2010) Model independence in multi-model ensemble prediction. Aust Meteorol Oceanogr J 59(1SP):3–6. https://doi.org/10.22499/2.5901.002

Abramowitz G, Gupta H (2008) Toward a model space and model independence metric. Geophys Res Lett. https://doi.org/10.1029/2007GL032834

Abramowitz G, Herger N, Gutmann E, Hammerling D, Knutti R, Leduc M, Lorenz R, Pincus R, Schmidt GA (2018) Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. Earth Syst Dyn Discuss. https://doi.org/10.5194/esd-2018-51

Allen RG (1998) Crop evapotranspiration: guidelines for computing crop water requirements. FAO irrigation and drainage paper, vol. 56. FAO, Rome

Arsenault R, Gatien P, Renaud B, Brissette F, Martel J-L (2015) A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J Hydrol 529:754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001

Asseng S, Ewert F, Rosenzweig C, Jones JW, Hatfield JL, Ruane AC, Boote KJ, Thorburn PJ, Rötter RP, Cammarano D, Brisson N, Basso B, Martre P, Aggarwal PK, Angulo C, Bertuzzi P, Biernath C, Challinor AJ, Doltra J, Gayler S, Goldberg R, Grant R, Heng L, Hooker J, Hunt LA, Ingwersen J, Izaurralde RC, Kersebaum KC, Müller C, Naresh Kumar S, Nendel C, O'Leary G, Olesen JE, Osborne TM, Palosuo T, Priesack E, Ripoche D, Semenov MA, Shcherbak I, Steduto P, Stöckle C, Stratonovitch P, Streck T, Supit I, Tao F, Travasso M, Waha K, Wallach D, White JW, Williams JR, Wolf J (2013) Uncertainty in simulating wheat yields under climate change. Nat Clim Change 3(9):827–832. https://doi.org/10.1038/nclimate1916

Asseng S, Ewert F, Martre P, Rötter RP, Lobell DB, Cammarano D, Kimball BA, Ottman MJ, Wall GW, White JW, Reynolds MP, Alderman PD, Prasad PVV, Aggarwal PK, Anothai J, Basso B, Biernath C, Challinor AJ, deSanctis G, Doltra J, Fereres E, GarciaVila M, Gayler S, Hoogenboom G, Hunt LA, Izaurralde RC, Jabloun M, Jones CD, Kersebaum KC, Koehler A-K, Müller C, NareshKumar S, Nendel C, O'Leary G, Olesen JE, Palosuo T, Priesack E, EyshiRezaei E, Ruane AC, Semenov MA, Shcherbak I, Stöckle C, Stratonovitch P, Streck T, Supit I, Tao F, Thorburn PJ, Waha K, Wang E, Wallach D, Wolf J, Zhao Z, Zhu Y (2015) Rising temperatures reduce global wheat production. Nat Clim Change 5(2):143–147. https://doi.org/10.1038/NCLIMATE2470

Bennett A, Nijssen B, Ou G, Clark M, Nearing G (2019) Quantifying Process connectivity with transfer entropy in hydrologic models. Water Resour Res. https://doi.org/10.1029/2018WR024555

Biernath C, Gayler S, Bittner S, Klein C, Högy P, Fangmeier A, Priesack E (2011) Evaluating the ability of four crop models to predict different environmental impacts on spring wheat grown in open-top chambers. Eur J Agron 35(2):71–82. https://doi.org/10.1016/j.eja.2011.04.001

Bishop CH, Abramowitz G (2013) Climate model dependence and the replicate Earth paradigm. Clim Dyn 41(3–4):885–900. https://doi.org/10.1007/s00382-012-1610-y

Christiansen B (2018) Ensemble averaging and the curse of dimensionality. J Clim 31(4):1587–1596. https://doi.org/10.1175/JCLI-D-17-0197.1

Deza M, Deza E (2016) Encyclopedia of distances, 4th edn. Springer, Heidelberg

Diks CGH, Vrugt JA (2010) Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch Environ Res Risk Assess 24(6):809–820. https://doi.org/10.1007/s00477-010-0378-z

Doblas-Reyes FJ, Hagedorn R, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. Tellus A 57(3):234–252. https://doi.org/10.1111/j.1600-0870.2005.00104.x (**accessed 2018-07-24**)

Enemark T, Peeters LJ, Mallants D, Batelaan O, Valentine AP, 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

Evans JP, Ji F, Abramowitz G, Ekström M (2013) Optimally choosing small ensemble members to produce robust climate simulations. Environ Res Lett 8(4):044050. https://doi.org/10.1088/1748-9326/8/4/044050

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

Fritsch JM (2000) Model Consensus. Weather Forecast 15:571–582

Garthwaite PH, Mubwandarikwa E (2010) Selection of weights for weighted model averaging: prior weights for weighted model averaging. Aust N Zeal J Stat 52(4):363–382. https://doi.org/10.1111/j.1467-842X.2010.00589.x

Gayler S, Wang E, Priesack E, Schaaf T, Maidl F-X (2002) Modeling biomass growth, N-uptake and phenological development of potato crop. Geoderma 105(3):367–383. https://doi.org/10.1016/S0016-7061(01)00113-6

Gayler S, Ingwersen J, Priesack E, Wöhling T, Wulfmeyer V, Streck T (2013) Assessing the relevance of subsurface processes for the simulation of evapotranspiration and soil moisture dynamics with CLM3.5: comparison with field data and crop model simulations. Environ Earth Sci 69(2):415–427. https://doi.org/10.1007/s12665-013-2309-z

Georgakakos KP, Seo D-J, Gupta H, Schaake J, Butts MB (2004) Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. J Hydrol 298(1–4):222–241. https://doi.org/10.1016/j.jhydrol.2004.03.037

George EI (2010) Dilution priors: Compensating for model space redundancy. In: Institute of Mathematical Statistics Collections. Institute of Mathematical Statistics, Beachwood, Ohio, USA, pp 158–165. https://doi.org/10.1214/10-IMSCOLL611

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437

Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. Tellus A 57(3):219–233. https://doi.org/10.1111/j.1600-0870.2005.00103.x

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

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

Hutson JL, Wagenet RJ (1995) An Overview of LEACHM: A Process Based Model of Water and Solute Movement, Transformations, Plant Uptake and Chemical Reactions in the Unsaturated Zone. In: Loeppert RH, Schwab AP, Goldberg S (eds) Chemical equilibrium and reaction models. SSSA Special Publications, Soil Science Society of America and American Society of Agronomy, Madison, WI, USA, pp. 409–422. https://doi.org/10.2136/sssaspecpub42.c19

Jefferys WH, Berger JO (1992) Ockham's Razor and Bayesian analysis. American Scientist 80(1):64–72.http://www.jstor.org/stable/29774559

Johnsson H, Bergstrom L, Jansson P-E, Paustian K (1987) Simulated nitrogen dynamics and losses in a layered agricultural soil. Agric Ecosyst Environ 18(4):333–356. https://doi.org/10.1016/0167-8809(87)90099-5

Jones CA (1986) CERES-Maize; a simulation model of maize growth and development vol. 04; SB91. M2, J6

Knutti R, Sedláček J, Sanderson BM, Lorenz R, Fischer EM, Eyring V (2017) A climate model projection weighting scheme accounting for performance and interdependence: model projection weighting scheme. Geophys Res Lett. https://doi.org/10.1002/2016GL072012 (**accessed 2018-07-30**)

Krishnamurti TN, Kishtawal CM, Zhang Z, LaRow T, Bachiochi D, Williford E, Gadgil S, Surendran S (2000) Multimodel ensemble forecasts for weather and seasonal climate. J Clim

13(23):4196–4216. https://doi.org/10.1175/1520-0442(2000)0134196:MEFFWA2.0.CO;2 (**accessed 2018-07-24**)
Lever J, Krzywinski M, Altman N (2016) Model selection and overfitting. Nat Methods 13(9):703–704. https://doi.org/10.1038/nmeth.3968
Liu S, Maljovec D, Wang B, Bremer P-T, Pascucci V (2017) Visualizing high-dimensional data: advances in the past decade. IEEE Trans Vis Comput Graph 23(3):1249–1268. https://doi.org/10.1109/TVCG.2016.2640960
Makowski D (2017) A simple Bayesian method for adjusting ensemble of crop model outputs to yield observations. Eur J Agron 88:76–83. https://doi.org/10.1016/j.eja.2015.12.012
...Martre P, Wallach D, Asseng S, Ewert F, Jones JW, Rötter RP, Boote KJ, Ruane AC, Thorburn PJ, Cammarano D, Hatfield JL, Rosenzweig C, Aggarwal PK, Angulo C, Basso B, Bertuzzi P, Biernath C, Brisson N, Challinor AJ, Doltra J, Gayler S, Goldberg R, Grant RF, Heng L, Hooker J, Hunt LA, Ingwersen J, Izaurralde RC, Kersebaum KC, Müller C, Kumar SN, Nendel C, O'leary G, Olesen JE, Osborne TM, Palosuo T, Priesack E, Ripoche D, Semenov MA, Shcherbak I, Steduto P, Stöckle CO, Stratonovitch P, Streck T, Supit I, Tao F, Travasso M, Waha K, White JW, Wolf J (2015) Multimodel ensembles of wheat growth many models are better than one. Glob Change Biol 21(2):911–925. https://doi.org/10.1111/gcb.12768
Minka TP (2002) Bayesian model averaging is not model combination. Technical report
Nandi G, Sharma RK (2020) Data Science fundamentals and practical approaches: understand why data science is the next. BPB Publications, Delhi
Nearing GS, Gupta HV (2018) Ensembles vs. information theory: supporting science under uncertainty. Front Earth Sci 12(4):653–660. https://doi.org/10.1007/s11707-018-0709-9
Novoselova N, Wang J, Klawonn F (2015) Optimized leaf ordering with class labels for hierarchical clustering. J Bioinform Comput Biol 13(04):1550012. https://doi.org/10.1142/S0219720015500122
Palosuo T, Kersebaum KC, Angulo C, Hlavinka P, Moriondo M, Olesen JE, Patil RH, Ruget F, Rumbaur C, Takáč J, Trnka M, Bindi M, Çaldağ B, Ewert F, Ferrise R, Mirschel W, Şaylan L, Šiška B, Rötter R (2011) Simulation of winter wheat yield and its variability in different climates of Europe: a comparison of eight crop growth models. Eur J Agron 35(3):103–114. https://doi.org/10.1016/j.eja.2011.05.001
Priesack E (2006) Expert-N Dokumentation der Modellbibliothek: Zugl.: Göttingen, Univ., Habil.-Schr., 2006. FAM-Bericht, vol. 60. Hieronymus, München
Priesack E, Gayler S (2009) Agricultural crop models: concepts of resource acquisition and assimilate partitioning. In: Lüttge U, Beyschlag W, Büdel B, Francis D (eds) Progress in botany. Springer, Berlin, pp 195–222. https://doi.org/10.1007/978-3-540-68421-3_9
Rettie FM, Gayler S, Weber TKD, Tesfaye K, Streck T (2022) Climate change impact on wheat and maize growth in Ethiopia: a multi-model uncertainty analysis. PLoS One 17(1):1. https://doi.org/10.1371/journal.pone.0262951
Richards LA (1931) Capillary conduction of liquids through porous mediums. J Appl Phys 1(5):318–333
Richardson LF (1922) Weather prediction by numerical process. Cambridge University Press, Cambridge
Ritchie JT, Godwin D (1989) CERES Wheat 2.0. Publication Title: CERES Wheat 2.0
Ritchie JT, Godwin DC, Otter-Nacke S (1988) CERES-Wheat. A simulation model of wheat growth and development. University of Texas Press, Austin
Rizzo ML, Székely GJ (2016) Energy distance. Wiley Interdiscip Rev Comput Stat 8(1):27–38. https://doi.org/10.1002/wics.1375
Rosenzweig C, Jones JW, Hatfield JL, Ruane AC, Boote KJ, Thorburn P, Antle JM, Nelson GC, Porter C, Janssen S, Asseng S, Basso B, Ewert F, Wallach D, Baigorria G, Winter JM (2013) The Agricultural Model Intercomparison and Improvement Project (AgMIP): protocols and pilot studies. Agric For Meteorol 170:166–182. https://doi.org/10.1016/j.agrformet.2012.09.011
Sanderson BM, Knutti R, Caldwell P (2015) Addressing interdependency in a multimodel ensemble by interpolation of model properties. J Clim 28(13):5150–5170. https://doi.org/10.1175/JCLI-D-14-00361.1
Sanderson BM, Knutti R, Caldwell P (2015) A representative democracy to reduce interdependency in a multimodel ensemble. J Clim 28(13):5171–5194. https://doi.org/10.1175/JCLI-D-14-00362.1
Šmunek J, Huang K, van Genuchten MT (1998) The HYDRUS code for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media, version 6.0: Research Report No. 144. U.S. Salinity Laboratory, Riverside, California
Spitters CJT, van Keulen H, van Kraalingen DWG (1989) A simple and universal crop growth simulator: SUCROS87. In: Rabbinge R, Ward SA, van Laar HH (eds) Simulation and systems management in crop protection. Simulation monographs. Pudoc, Wageningen, The Netherlands, pp 147–181
Streck T, Weber TKD (2020) Analytical expressions for noncapillary soil water retention based on popular capillary retention models. Vadose Zone J 19:e20042. https://doi.org/10.1002/vzj2.20042
Székely GJ, Rizzo ML (2013) Energy statistics: a class of statistics based on distances. J Stat Plan Infer 143(8):1249–1272. https://doi.org/10.1016/j.jspi.2013.03.018
Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A Math Phys Eng Sci 365(1857):2053–2075. https://doi.org/10.1098/rsta.2007.2076
van Dam JC, Groenendijk P, Hendriks RFA, Kroes JG (2008) Advances of modeling water flow in variably saturated soils with SWAP. Vadose Zone J 7(2):640. https://doi.org/10.2136/vzj2007.0060
van Genuchten MT (1980) Closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci Soc Am J 44(5):892–898
van Laar HH, Goudriaan J, Keulen H (1997) Sucros97: Simulation of Crop Growth for Potential and Water-limited Situations. Service Agricultural Sciences (DLO), Wageningen, The Netherlands. Backup Publisher: Res. Inst. for Agrobiol. and Soil-Fertility and The C.T. de Wit Graduate Schl. for Prod. Ecol
Van Laar H, Goudriaan J, Van Keulen H (1992) Simulation of crop growth for potential and water-limited production situations: as applied to spring wheat. Technical report, CABO-DLO
Vehtari A, Ojanen J (2012) A survey of Bayesian predictive methods for model assessment, selection and comparison. Stat Surv 6:142–228. https://doi.org/10.1214/12-SS102
Vogel RM, Sankarasubramanian A (2003) Validation of a watershed model without calibration. Water Resour Res. https://doi.org/10.1029/2002WR001940
Wallach D (2011) Crop model calibration: a statistical perspective. Agron J 103(4):1144–1151. https://doi.org/10.2134/agronj2010.0432
Wallach D, Martre P, Liu B, Asseng S, Ewert F, Thorburn PJ, van Ittersum M, Aggarwal PK, Ahmed M, Basso B, Biernath C, Cammarano D, Challinor AJ, De Sanctis G, Dumont B, Eyshi Rezaei E, Fereres E, Fitzgerald GJ, Gao Y, Garcia-Vila M, Gayler S, Girousse C, Hoogenboom G, Horan H, Izaurralde RC, Jones CD, Kassie BT, Kersebaum KC, Klein C, Koehler AK, Maiorano A, Minoli S, Müller C, Naresh Kumar S, Nendel C, O'Leary GJ, Palosuo T, Priesack E, Ripoche D, Rötter RP, Semenov MA, Stöckle C, Stratonovitch P, Streck T, Supit I, Tao F, Wolf J, Zhang Z (2018) Multi-model ensembles improve predictions of

crop–environment–management interactions. Glob Change Biol. https://doi.org/10.1111/gcb.14411

Wallach D, Palosuo T, Thorburn P, Hochman Z, Gourdain E, Andrianasolo F, Asseng S, Basso B, Buis S, Crout N, Dibari C, Dumont B, Ferrise R, Gaiser T, Garcia C, Gayler S, Ghahramani A, Hiremath S, Hoek S, Horan H, Hoogenboom G, Huang M, Jabloun M, Jansson PE, Jing Q, Justes E, Kersebaum KC, Klosterhalfen A, Launay M, Lewan E, Luo Q, Maestrini B, Mielenz H, Moriondo M, Zadeh HN, Padovan G, Olesen JE, Poyda A, Priesack E, Pullens JWM, Qian B, Schütze N, Shelia V, Souissi A, Specka X, Srivastava AK, Stella T, Streck T, Trombi G, Wallor E, Wang J, Weber TKD, Weihermüller L, de Wit A, Wöhling T, Xiao L, Zhao C, Zhu Y, Seidel SJ (2020) The chaos in calibrating crop models. Plant Biol. https://doi.org/10.1101/2020.09.12.294744

Wang E (1997) Development of a generic process-oriented model for simulation of crop growth. Ökologie (Munich, Germany). Utz, Wissenschaft, Munich

Wang E, Engel T (1998) Simulation of phenological development of wheat crops. Agric Syst 58(1):1–24. https://doi.org/10.1016/S0308-521X(98)00028-6

Wang E, Engel T (2000) SPASS: a generic process-oriented crop model with versatile windows interfaces. Environ Model Softw 15(2):179–188. https://doi.org/10.1016/S1364-8152(99)00033-X

Weber TKD, Durner W, Streck T, Diamantopoulos E (2019) A modular framework for modelling unsaturated soil hydraulic properties over the full moisture range. Water Resour Res. https://doi.org/10.1029/2018WR024584

Weber TKD, Finkel M, Conceição Gonçalves M, Vereecken H, Diamantopoulos E (2020) Pedotransfer function for the Brunswick soil hydraulic property model and comparison to the van Genuchten-Mualem model. Water Resour Res. https://doi.org/10.1029/2019WR026820

Weber TKD, Ingwersen J, Högy P, Poyda A, Wizemann H-D, Demyan MS, Bohm K, Eshonkulov R, Gayler S, Kremer P, Laub M, Nkwain YF, Troost C, Witte I, Cadisch G, Müller T, Fangmeier A, Wulfmeyer V, Streck T (2021) Multi-site, multi-crop measurements in the soil-vegetation-atmosphere continuum: a comprehensive dataset from two climatically contrasting regions in South West Germany for the period 2009–2018. Earth Syst Sci Data Discuss 2021:1–32. https://doi.org/10.5194/essd-2020-396

Weigel AP, Liniger MA, Appenzeller C (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134(630):241–260. https://doi.org/10.1002/qj.210 (**accessed 2018-07-24**)

Winter CL, Nychka D (2010) Forecasting skill of model averages. Stoch Environ Res Risk Assess 24(5):633–638. https://doi.org/10.1007/s00477-009-0350-y

Wöhling T, Geiges A, Nowak W, Gayler S, Högy P, Wizemann HD (2013) Towards optimizing experiments for maximum-confidence model selection between different soil–plant models. Procedia Environ Sci 19:514–523. https://doi.org/10.1016/j.proenv.2013.06.058

Wöhling T, Schöniger A, Gayler S, Nowak W (2015) Bayesian model averaging to explore the worth of data for soil–plant model selection and prediction. Water Resour Res 51(4):2825–2846. https://doi.org/10.1002/2014WR016292

Xu R, Wunsch D (2008) Clustering. IEEE Press Series on Computational Intelligence. Wiley, New Jersey

Yao Y, Vehtari A, Simpson D, Gelman A (2018) Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Anal 13(3):917–1007. https://doi.org/10.1214/17-BA1091

Yun K, Hsiao J, Jung M-P, Choi I-T, Glenn DM, Shim K-M, Kim S-H (2017) Can a multi-model ensemble improve phenology predictions for climate change studies? Ecol Model 362:54–64. https://doi.org/10.1016/j.ecolmodel.2017.08.003

Ziel F, Berk K (2019) Multivariate forecasting evaluation: on sensitive and strictly proper scoring rules. arXiv:1910.07325 [econ, stat]

Bibliography

G. Abramowitz. Model independence in multi-model ensemble prediction. *Australian Meteorological and Oceanographic Journal*, 59(1SP):3–6, 2010. ISSN 1836716X. doi: 10.22499/2.5901.002.

G. Abramowitz and C. H. Bishop. Climate Model Dependence and the Ensemble Dependence Transformation of CMIP Projections. *Journal of Climate*, 28(6):2332–2348, Mar. 2015. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-14-00364.1.

G. Abramowitz and H. Gupta. Toward a model space and model independence metric. *Geophysical Research Letters*, 35(5), Mar. 2008. ISSN 0094-8276. doi: 10.1029/2007GL032834.

G. Abramowitz, N. Herger, E. Gutmann, D. Hammerling, R. Knutti, M. Leduc, R. Lorenz, R. Pincus, and G. A. Schmidt. Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth System Dynamics Discussions*, pages 1–20, July 2018. ISSN 2190-4995. doi: 10.5194/esd-2018-51.

N. K. Ajami, Q. Duan, and S. Sorooshian. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, 43(1), Jan. 2007. ISSN 00431397. doi: 10.1029/2005WR004745.

E. Alpaydin. *Introduction to Machine Learning*. Adaptive computation and machine learning. MIT Press, 2004. ISBN 978-0-262-01211-9.

R. Arsenault, P. Gatien, B. Renaud, F. Brissette, and J.-L. Martel. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. *Journal of Hydrology*, 529:754–767, Oct. 2015. ISSN 00221694. doi: 10.1016/j.jhydrol.2015.09.001.

B. M. Ayyub and G. J. Klir. *Uncertainty Modeling and Analysis in Engineering and the Sciences*. Chapman and Hall/CRC, 0 edition, May 2006. ISBN 978-0-429-19184-8. doi: 10.1201/9781420011456.

J. E. M. Baartman, L. A. Melsen, D. Moore, and M. J. v. d. Ploeg. On the complexity of model complexity: Viewpoints across the geosciences. *Catena*, 186:104261, 2020. ISSN 0341-8162. doi: https://doi.org/10.1016/j.catena.2019.104261.

G. J. Babu. Resampling Methods for Model Fitting and Model Selection. *Journal of Biopharmaceutical Statistics*, 21(6):1177–1186, Nov. 2011. ISSN 1054-3406, 1520-5711. doi: 10.1080/10543406.2011.607749.

A. Bennett, B. Nijssen, G. Ou, M. Clark, and G. Nearing. Quantifying Process Connectivity With Transfer Entropy in Hydrologic Models. *Water Resources Research*, June 2019. ISSN 0043-1397, 1944-7973. doi: 10.1029/2018WR024555.

J. Bernardo and A. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 978-0-470-31771-6.

J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Bayesian Model Averaging and Model Search Strategies. 1999.

C. H. Bishop and G. Abramowitz. Climate model dependence and the replicate Earth paradigm. *Climate Dynamics*, 41(3-4):885–900, Aug. 2013. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-012-1610-y.

C. M. Bishop. Neural Networks for Pattern Recognition. page 498, 1995.

G. Brunetti, J. Šimůnek, D. Glöckler, and C. Stumpp. Handling model complexity with parsimony: Numerical analysis of the nitrogen turnover in a controlled aquifer model setup. *Journal of Hydrology*, 584:124681, May 2020. ISSN 00221694. doi: 10.1016/j.jhydrol.2020.124681.

M. D. Buhmann. *Radial Basis Functions: Theory and Implementations*. Cambridge University Press, 1 edition, July 2003. ISBN 978-0-521-63338-3 978-0-521-10133-2 978-0-511-54324-1. doi: 10.1017/CBO9780511543241.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York, 2nd ed edition, 2002. ISBN 978-0-387-95364-9. OCLC: ocm48557578.

G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46 (2):255–308, Jan. 2009. ISSN 0273-0979. doi: 10.1090/S0273-0979-09-01249-X.

F. Chazal and B. Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *arXiv:1710.04019 [cs, math, stat]*, Oct. 2017.

S. Cheng and K. Mueller. The Data Context Map: Fusing Data and Attributes into a Unified Display. *IEEE Transactions on Visualization and Computer Graphics*, 22 (1):121–130, Jan. 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467552.

B. Christiansen. Ensemble Averaging and the Curse of Dimensionality. *Journal of Climate*, 31(4):1587–1596, Feb. 2018. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-17-0197.1.

M. Deza and E. Deza. *Encyclopedia of distances*. Springer, Heidelberg ; New York, fourth edition edition, 2016. ISBN 978-3-662-52843-3.

C. G. H. Diks and J. A. Vrugt. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stochastic Environmental Research and Risk Assessment*, 24(6):809–820, Aug. 2010. ISSN 1436-3240, 1436-3259. doi: 10.1007/s00477-010-0378-z.

F. J. Doblas-Reyes, R. Hagedorn, and T. N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. *Tellus A*, 57(3):234–252, May 2005. ISSN 0280-6495, 1600-0870. doi: 10.1111/j.1600-0870.2005.00104.x.

T. Enemark, L. J. Peeters, D. Mallants, O. Batelaan, A. P. Valentine, and M. Sambridge. Hydrogeological Bayesian Hypothesis Testing through Trans-Dimensional Sampling of a Stochastic Water Balance Model. *Water*, 11(7):1463, July 2019. ISSN 2073-4441. doi: 10.3390/w11071463.

J. P. Evans, F. Ji, G. Abramowitz, and M. Ekström. Optimally choosing small ensemble members to produce robust climate simulations. *Environmental Research Letters*, 8 (4):044050, Dec. 2013. ISSN 1748-9326. doi: 10.1088/1748-9326/8/4/044050.

B. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 978-0-470-74991-3.

T. P. Ferré. Revisiting the Relationship Between Data, Models, and Decision-Making. *Groundwater*, 55(5):604–614, Sept. 2017. ISSN 0017467X. doi: 10.1111/gwat.12574.

J. M. Fritsch. Model Consensus. *Weather and Forecasting*, 15:571–582, 2000.

P. H. Garthwaite and E. Mubwandarikwa. Selection of weights for weighted model averaging: Prior weights for model averaging. *Australian & New Zealand Journal of Statistics*, 52(4):363–382, Dec. 2010. ISSN 13691473. doi: 10.1111/j.1467-842X.2010.00589.x.

S. Gayler, E. Wang, E. Priesack, T. Schaaf, and F.-X. Maidl. Modeling biomass growth, N-uptake and phenological development of potato crop. *Geoderma*, 105(3):367–383, 2002. ISSN 0016-7061. doi: https://doi.org/10.1016/S0016-7061(01)00113-6.

S. Geman, E. Bienenstock, and R. Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 1992.

K. P. Georgakakos, D.-J. Seo, H. Gupta, J. Schaake, and M. B. Butts. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology*, 298(1-4):222–241, Oct. 2004. ISSN 00221694. doi: 10.1016/j.jhydrol.2004.03.037.

E. I. George. Dilution priors: Compensating for model space redundancy. In *Institute of Mathematical Statistics Collections*, pages 158–165. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2010. ISBN 978-0-940600-79-9. doi: 10.1214/10-IMSCOLL611.

W. M. Getz, C. R. Marshall, C. J. Carlson, L. Giuggioli, S. J. Ryan, S. S. Romañach, C. Boettiger, S. D. Chamberlain, L. Larsen, P. D'Odorico, and D. O'Sullivan. Making ecological models adequate. *Ecology Letters*, 21(2):153–166, Feb. 2018. ISSN 1461023X. doi: 10.1111/ele.12893.

T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, Mar. 2007. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214506000001437.

H. V. Gupta, M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye. Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, 48 (8), Aug. 2012. ISSN 00431397. doi: 10.1029/2011WR011044.

A. Guthke. Defensible Model Complexity: A Call for Data-Based and Goal-Oriented Model Choice. *Ground Water*, 55(5):646–650, 2017. ISSN 0017-467X. doi: 10.1111/gwat.12554.

J. Görtler, T. Spinner, D. Streeb, D. Weiskopf, and O. Deussen. Uncertainty-Aware Principal Component Analysis. *arXiv:1905.01127 [cs, stat]*, May 2019. arXiv: 1905.01127.

R. Hagedorn, F. J. Doblas-Reyes, and T. N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus A*, 57(3):219–233, May 2005. ISSN 0280-6495, 1600-0870. doi: 10.1111/j.1600-0870.2005.00103.x.

E. Hellinger. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

J. Hommel, E. Lauchnor, A. Phillips, R. Gerlach, A. B. Cunningham, R. Helmig, A. Ebigbo, and H. Class. A revised model for microbially induced calcite precipitation: Improvements and new insights based on recent experiments. *Water Resources Research*, 51(5):3695–3715, 2015. Publisher: Wiley Online Library.

J. Hommel, A. Ebigbo, R. Gerlach, A. B. Cunningham, R. Helmig, and H. Class. Finding a Balance between Accuracy and Effort For Modeling Biomineralization. *Energy Procedia*, 97:379–386, Nov. 2016. ISSN 18766102. doi: 10.1016/j.egypro.2016.10.028.

M. Höge. *Bayesian Multi-Model Frameworks Properly Addressing Conceptual Uncertainty in Applied Modelling*. PhD thesis, Tübingen, Tübingen, 2019.

M. Höge, T. Wöhling, and W. Nowak. A Primer for Model Selection: The Decisive Role of Model Complexity. *Water Resources Research*, 54(3):1688–1715, Mar. 2018. ISSN 00431397. doi: 10.1002/2017WR021902.

M. Höge, A. Guthke, and W. Nowak. The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, 572:96–107, 2019. ISSN 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2019.01.072.

M. Höge, A. Guthke, and W. Nowak. Bayesian Model Weighting: The Many Faces of Model Averaging. *Water*, 12(2):309, Jan. 2020. ISSN 2073-4441. doi: 10.3390/w12020309.

H. Jeffreys. *Theory of probability, Clarendon*. Oxford, 1961.

C. Johnson. Top scientific visualization research problems. *IEEE Computer Graphics and Applications*, 24(4):13–17, July 2004. ISSN 0272-1716, 1558-1756. doi: 10.1109/MCG.2004.20.

R. E. Kass and A. E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995. ISSN 0162-1459. doi: 10.1080/01621459.1995.10476572.

R. Knutti, R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl. Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate*, 23(10):2739–2758, May 2010. ISSN 1520-0442, 0894-8755. doi: 10.1175/2009JCLI3361.1.

R. Knutti, J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring. A climate model projection weighting scheme accounting for performance and interdependence: Model Projection Weighting Scheme. *Geophysical Research Letters*, 2017. ISSN 00948276. doi: 10.1002/2016GL072012.

D. G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951. Publisher: Sabinet.

T. N. Krishnamurti, C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran. Multimodel Ensemble Forecasts for Weather and Seasonal Climate. *Journal of Climate*, 13(23):4196–4216, Dec. 2000. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(2000)013⟨4196:MEFFWA⟩2.0.CO;2.

J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964. ISSN 1860-0980. doi: 10.1007/BF02289565.

M. Köppel, F. Franzelin, I. Kröker, S. Oladyshkin, G. Santin, D. Wittwar, A. Barth, B. Haasdonk, W. Nowak, D. Pflüger, and C. Rohde. Comparison of data-driven uncertainty quantification methods for a carbon dioxide storage benchmark scenario.

Computational Geosciences, 23(2):339–354, Apr. 2019. ISSN 1420-0597, 1573-1499. doi: 10.1007/s10596-018-9785-x.

J. Lever, M. Krzywinski, and N. Altman. Model selection and overfitting. *Nature Methods*, 13(9):703–704, Sept. 2016. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3968.

S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing High-Dimensional Data: Advances in the Past Decade. *IEEE Transactions on Visualization and Computer Graphics*, 23(3):1249–1268, Mar. 2017. ISSN 1077-2626. doi: 10.1109/TVCG.2016.2640960.

R. Lorenz, N. Herger, J. Sedláček, V. Eyring, E. M. Fischer, and R. Knutti. Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America. *Journal of Geophysical Research: Atmospheres*, 123 (9):4509–4526, May 2018. ISSN 2169897X. doi: 10.1029/2017JD027992.

M. Loschko, T. Wöhling, D. L. Rudolph, and O. A. Cirpka. Cumulative relative reactivity: A concept for modeling aquifer-scale reactive transport. *Water Resources Research*, 52(10):8117–8137, Oct. 2016. ISSN 00431397. doi: 10.1002/2016WR019080.

D. Lu, M. Ye, and G. P. Curtis. Maximum likelihood Bayesian model averaging and its predictive analysis for groundwater reactive transport models. *Journal of Hydrology*, 529:1859–1873, Oct. 2015. ISSN 00221694. doi: 10.1016/j.jhydrol.2015.07.029.

D. J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.415.

L. N. Marelli, S. and B. Sudret. UQLab user manual – Polynomial chaos expansions. Technical report, Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland, 2021.

G. A. Martos Venturini. Statistical distances and probability metrics for multivariate data, ensembles and probability distributions. 2015.

P. Martre, D. Wallach, S. Asseng, F. Ewert, J. W. Jones, R. P. Rötter, K. J. Boote, A. C. Ruane, P. J. Thorburn, D. Cammarano, J. L. Hatfield, C. Rosenzweig, P. K. Aggarwal, C. Angulo, B. Basso, P. Bertuzzi, C. Biernath, N. Brisson, A. J. Challinor, J. Doltra, S. Gayler, R. Goldberg, R. F. Grant, L. Heng, J. Hooker, L. A. Hunt,

J. Ingwersen, R. C. Izaurralde, K. C. Kersebaum, C. Müller, S. N. Kumar, C. Nendel, G. O'leary, J. E. Olesen, T. M. Osborne, T. Palosuo, E. Priesack, D. Ripoche, M. A. Semenov, I. Shcherbak, P. Steduto, C. O. Stöckle, P. Stratonovitch, T. Streck, I. Supit, F. Tao, M. Travasso, K. Waha, J. W. White, and J. Wolf. Multimodel ensembles of wheat growth: many models are better than one. *Global Change Biology*, 21(2):911–925, Feb. 2015. ISSN 13541013. doi: 10.1111/gcb.12768.

G. Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, 1963. ISSN 0361-0128. doi: 10.2113/gsecongeo.58.8.1246.

T. P. Minka. Bayesian model averaging is not model combination. Technical report, 2002.

F. Mohammadi, R. Kopmann, A. Guthke, S. Oladyshkin, and W. Nowak. Bayesian selection of hydro-morphodynamic models under computational time constraints. *Advances in Water Resources*, 117:53–64, July 2018. ISSN 03091708. doi: 10.1016/j.advwatres.2018.05.007.

K. Monteith, J. L. Carroll, K. Seppi, and T. Martinez. Turning Bayesian model averaging into Bayesian model combination. pages 2657–2663. IEEE, July 2011. ISBN 978-1-4244-9635-8. doi: 10.1109/IJCNN.2011.6033566.

K. Mosler. Depth Statistics. In C. Becker, R. Fried, and S. Kuhnt, editors, *Robustness and Complex Data Structures*, pages 17–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-35493-9 978-3-642-35494-6. doi: 10.1007/978-3-642-35494-6_2.

G. Nandi and R. Sharma. *Data Science Fundamentals and Practical Approaches: Understand Why Data Science Is the Next*. BPB Publications, Delhi, 2020. ISBN 978-93-89845-66-2.

G. S. Nearing and H. V. Gupta. Ensembles vs. information theory: supporting science under uncertainty. *Frontiers of Earth Science*, 12(4):653–660, Dec. 2018. ISSN 2095-0195, 2095-0209. doi: 10.1007/s11707-018-0709-9.

W. Neuman. A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites. page 311, 2003.

S. Oladyshkin. *aPC Matlab Toolbox: Data-driven Arbitrary Polynomial Chaos, Matlab Central File Exchange*. Retrieved October 15, 2020, 2020.

S. Oladyshkin and W. Nowak. Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. *Reliability Engineering & System Safety*, 106: 179–190, Oct. 2012. ISSN 09518320. doi: 10.1016/j.ress.2012.05.002.

S. Oladyshkin, F. P. J. d. Barros, and W. Nowak. Global sensitivity analysis: A flexible and efficient framework with an example from stochastic hydrogeology. *Advances in Water Resources*, 37:10–22, 2012. ISSN 0309-1708. doi: https://doi.org/10.1016/j.advwatres.2011.11.001.

S. Oladyshkin, H. Class, and W. Nowak. 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. *Computational Geosciences*, 17(4):671–687, 2013. Publisher: Springer.

T. Palosuo, K. C. Kersebaum, C. Angulo, P. Hlavinka, M. Moriondo, J. E. Olesen, R. H. Patil, F. Ruget, C. Rumbaur, J. Takáč, M. Trnka, M. Bindi, B. Çaldağ, F. Ewert, R. Ferrise, W. Mirschel, L. Şaylan, B. Šiška, and R. Rötter. Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. *European Journal of Agronomy*, 35(3):103–114, Oct. 2011. ISSN 11610301. doi: 10.1016/j.eja.2011.05.001.

E. Priesack. *Expert-N-Dokumentation der Modellbibliothek*. FAM-Bericht. Hieronymus, 2006. ISBN 978-3-89791-362-2.

A. E. Raftery. Bayesian Model Selection in Social Research. *Sociological Methodology*, 25, 1995.

J. C. Refsgaard, S. Christensen, T. O. Sonnenborg, D. Seifert, A. L. Højberg, and L. Troldborg. Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Advances in Water Resources*, 36:36–50, Feb. 2012. ISSN 03091708. doi: 10.1016/j.advwatres.2011.04.006.

S. L. Rinderknecht, M. E. Borsuk, and P. Reichert. Bridging uncertain and ambiguous knowledge with imprecise probabilities. *Environmental Modelling & Software*, 36: 122–130, Oct. 2012. ISSN 13648152. doi: 10.1016/j.envsoft.2011.07.022.

J. T. Ritchie, D. C. Godwin, and S. Otter-Nacke. *CERES-Wheat. A Simulation Model of Wheat Growth and Development*. University of Texas Press, Austin, 1988.

M. L. Rizzo and G. J. Székely. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38, Jan. 2016. ISSN 19395108. doi: 10.1002/wics.1375.

R. Rojas, L. Feyen, and A. Dassargues. Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research*, 44(12), Dec. 2008. ISSN 00431397. doi: 10.1029/2008WR006908.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. [Design and Analysis of Computer Experiments]: Rejoinder. *Statistical Science*, 4(4):433–435, Nov. 1989. ISSN 0883-4237. doi: 10.1214/ss/1177012420.

B. M. Sanderson, R. Knutti, and P. Caldwell. Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties. *Journal of Climate*, 28(13):5150–5170, July 2015a. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-14-00361.1.

B. M. Sanderson, R. Knutti, and P. Caldwell. A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. *Journal of Climate*, 28(13):5171–5194, July 2015b. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-14-00362.1.

A. Sanz-Prat, C. Lu, M. Finkel, and O. A. Cirpka. On the validity of travel-time based nonlinear bioreactive transport models in steady-state flow. *Journal of Contaminant Hydrology*, 175-176:26–43, Apr. 2015. ISSN 01697722. doi: 10.1016/j.jconhyd.2015.02.003.

S. Scheurer, A. Schafer, R. Silva, F. Mohammadi, J. Hommel, S. Oladyshkin, B. Flemisch, and W. Nowak. Surrogate-based Bayesian Comparison of 2 Computationally Expensive Models: Application to 3 Microbially Induced Calcite Precipitation. page 37, 2021.

A. Schäfer Rodrigues Silva, T. K. D. Weber, S. Gayler, A. Guthke, M. Höge, W. Nowak, and T. Streck. Diagnosing similarities in probabilistic multi-model ensembles: an application to soil–plant-growth-modeling. *Modeling Earth Systems and Environment*,

June 2022. ISSN 2363-6203, 2363-6211. doi: 10.1007/s40808-022-01427-1. URL https://link.springer.com/10.1007/s40808-022-01427-1.

A. Schäfer Rodrigues Silva, A. Guthke, M. Höge, O. A. Cirpka, and W. Nowak. Strategies for Simplifying Reactive Transport Models: A Bayesian Model Comparison. *Water Resources Research*, 56(11), Nov. 2020. ISSN 0043-1397, 1944-7973. doi: 10.1029/2020WR028100.

A. Schöniger, T. Wöhling, L. Samaniego, and W. Nowak. Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research*, 50(12):9484–9513, Dec. 2014. ISSN 00431397. doi: 10.1002/2014WR016062.

A. Schöniger, W. A. Illman, T. Wöhling, and W. Nowak. Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. *Journal of Hydrology*, 531:96–110, Dec. 2015. ISSN 00221694. doi: 10.1016/j.jhydrol.2015.07.047.

S. K. Singh and A. Bárdossy. Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, 38:81–91, Mar. 2012. ISSN 03091708. doi: 10.1016/j.advwatres.2011.12.006.

C. Tebaldi and R. Knutti. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857):2053–2075, Aug. 2007. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2007.2076.

J. Tukey. Mathematics and picturing data. In *Canadian Math. Congress. MR0426989*, 1975.

H. H. van Laar, J. Goudriaan, and H. Keulen. *Sucros97: Simulation of crop growth for potential and water-limited situations*. Service Agricultural Sciences (DLO), Wageningen, The Netherlands, 1997. Backup Publisher: Res. Inst. for Agrobiol. and Soil-Fertility and The C.T. de Wit Graduate Schl. for Prod. Ecol.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer New York, New York, NY, 2000. ISBN 978-1-4419-3160-3 978-1-4757-3264-1. doi: 10.1007/978-1-4757-3264-1.

A. Vehtari and J. Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6(0):142–228, 2012. ISSN 1935-7516. doi: 10.1214/12-SS102.

R. M. Vogel and A. Sankarasubramanian. Validation of a watershed model without calibration. *Water Resources Research*, 39(10), Oct. 2003. ISSN 00431397. doi: 10.1029/2002WR001940.

D. Wallach. Crop Model Calibration: A Statistical Perspective. *Agronomy Journal*, 103(4):1144–1151, July 2011. ISSN 00021962. doi: 10.2134/agronj2010.0432.

D. Wallach, P. Martre, B. Liu, S. Asseng, F. Ewert, P. J. Thorburn, M. van Ittersum, P. K. Aggarwal, M. Ahmed, B. Basso, C. Biernath, D. Cammarano, A. J. Challinor, G. De Sanctis, B. Dumont, E. Eyshi Rezaei, E. Fereres, G. J. Fitzgerald, Y. Gao, M. Garcia-Vila, S. Gayler, C. Girousse, G. Hoogenboom, H. Horan, R. C. Izaurralde, C. D. Jones, B. T. Kassie, K. C. Kersebaum, C. Klein, A. K. Koehler, A. Maiorano, S. Minoli, C. Müller, S. Naresh Kumar, C. Nendel, G. J. O'Leary, T. Palosuo, E. Priesack, D. Ripoche, R. P. Rötter, M. A. Semenov, C. Stöckle, P. Stratonovitch, T. Streck, I. Supit, F. Tao, J. Wolf, and Z. Zhang. Multi-model ensembles improve predictions of crop-environment-management interactions. *Global Change Biology*, July 2018. ISSN 13541013. doi: 10.1111/gcb.14411.

D. Wallach, T. Palosuo, P. Thorburn, Z. Hochman, E. Gourdain, F. Andrianasolo, S. Asseng, B. Basso, S. Buis, N. Crout, C. Dibari, B. Dumont, R. Ferrise, T. Gaiser, C. Garcia, S. Gayler, A. Ghahramani, S. Hiremath, S. Hoek, H. Horan, G. Hoogenboom, M. Huang, M. Jabloun, P.-E. Jansson, Q. Jing, E. Justes, K. C. Kersebaum, A. Klosterhalfen, M. Launay, E. Lewan, Q. Luo, B. Maestrini, H. Mielenz, M. Moriondo, H. N. Zadeh, G. Padovan, J. E. Olesen, A. Poyda, E. Priesack, J. W. M. Pullens, B. Qian, N. Schütze, V. Shelia, A. Souissi, X. Specka, A. K. Srivastava, T. Stella, T. Streck, G. Trombi, E. Wallor, J. Wang, T. K. Weber, L. Weihermüller, A. de Wit, T. Wöhling, L. Xiao, C. Zhao, Y. Zhu, and S. J. Seidel. The chaos in calibrating crop models. preprint, Plant Biology, Sept. 2020.

E. Wang and T. Engel. SPASS: a generic process-oriented crop model with versatile windows interfaces. *Environmental Modelling & Software*, 15(2):179–188, 2000. ISSN 13648152. doi: 10.1016/S1364-8152(99)00033-X.

L. Wasserman. Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, 2000.

A. P. Weigel, M. A. Liniger, and C. Appenzeller. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, 134(630):241–260, Jan. 2008. ISSN 00359009, 1477870X. doi: 10.1002/qj.210.

N. Wiener. The Homogeneous Chaos. *American Journal of Mathematics*, 60(4):897, Oct. 1938. ISSN 00029327. doi: 10.2307/2371268.

C. L. Winter and D. Nychka. Forecasting skill of model averages. *Stochastic Environmental Research and Risk Assessment*, 24(5):633–638, July 2010. ISSN 1436-3240, 1436-3259. doi: 10.1007/s00477-009-0350-y.

T. Wöhling and J. A. Vrugt. Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models. *Water Resources Research*, 44(12), Dec. 2008. ISSN 00431397. doi: 10.1029/2008WR007154.

T. Wöhling, A. Schöniger, S. Gayler, and W. Nowak. Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction: Bayesian model averaging to explore the worth of data. *Water Resources Research*, 51(4): 2825–2846, Apr. 2015. ISSN 00431397. doi: 10.1002/2014WR016292.

D. Xiu. *Numerical methods for stochastic computations: a spectral method approach*. Princeton University Press, Princeton, N.J, 2010. ISBN 978-0-691-14212-8. OCLC: ocn466341417.

R. Xu and D. Wunsch. *Clustering*. IEEE Press Series on Computational Intelligence. Wiley, New Jersey, 2008. ISBN 978-0-470-38278-3.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, 13(3):917–1007, Sept. 2018. ISSN 1936-0975. doi: 10.1214/17-BA1091.

F. Ziel and K. Berk. Multivariate Forecasting Evaluation: On Sensitive and Strictly Proper Scoring Rules. *arXiv:1910.07325 [econ, stat]*, Oct. 2019. arXiv: 1910.07325.

Institut für Wasser- und Umweltsystemmodellierung
Universität Stuttgart

Pfaffenwaldring 61
70569 Stuttgart (Vaihingen)
Telefon (0711) 685 - 60156
Telefax (0711) 685 - 51073
E-Mail: iws@iws.uni-stuttgart.de
http://www.iws.uni-stuttgart.de

Verzeichnis der Mitteilungshefte

1 Röhnisch, Arthur: *Die Bemühungen um eine Wasserbauliche Versuchsanstalt an der Technischen Hochschule Stuttgart*, und
Fattah Abouleid, Abdel: *Beitrag zur Berechnung einer in lockeren Sand gerammten, zweifach verankerten Spundwand*, 1963

2 Marotz, Günter: *Beitrag zur Frage der Standfestigkeit von dichten Asphaltbelägen im Großwasserbau*, 1964

3 Gurr, Siegfried: *Beitrag zur Berechnung zusammengesetzter ebener Flächentrag-werke unter besonderer Berücksichtigung ebener Stauwände, mit Hilfe von Rand-wert- und Lastwertmatrizen*, 1965

4 Plica, Peter: *Ein Beitrag zur Anwendung von Schalenkonstruktionen im Stahlwasserbau*, und
Petrikat, Kurt: *Möglichkeiten und Grenzen des wasserbaulichen Versuchswesens*, 1966

5 Plate, Erich: *Beitrag zur Bestimmung der Windgeschwindigkeitsverteilung in der durch eine Wand gestörten bodennahen Luftschicht*, und
Röhnisch, Arthur; Marotz, Günter: *Neue Baustoffe und Bauausführungen für den Schutz der Böschungen und der Sohle von Kanälen, Flüssen und Häfen; Gestehungskosten und jeweilige Vorteile*, sowie
Unny, T.E.: *Schwingungsuntersuchungen am Kegelstrahlschieber*, 1967

6 Seiler, Erich: *Die Ermittlung des Anlagenwertes der bundeseigenen Binnenschiffahrtsstraßen und Talsperren und des Anteils der Binnenschiffahrt an diesem Wert*, 1967

7 *Sonderheft anläßlich des 65. Geburtstages von Prof. Arthur Röhnisch mit Beiträgen von* Benk, Dieter; Breitling, J.; Gurr, Siegfried; Haberhauer, Robert; Honekamp, Hermann; Kuz, Klaus Dieter; Marotz, Günter; Mayer-Vorfelder, Hans-Jörg; Miller, Rudolf; Plate, Erich J.; Radomski, Helge; Schwarz, Helmut; Vollmer, Ernst; Wildenhahn, Eberhard; 1967

8 Jumikis, Alfred: *Beitrag zur experimentellen Untersuchung des Wassernachschubs in einem gefrierenden Boden und die Beurteilung der Ergebnisse*, 1968

9 Marotz, Günter: *Technische Grundlagen einer Wasserspeicherung im natürlichen Untergrund*, 1968

10 Radomski, Helge: *Untersuchungen über den Einfluß der Querschnittsform wellenförmiger Spundwände auf die statischen und rammtechnischen Eigenschaften*, 1968

11 Schwarz, Helmut: *Die Grenztragfähigkeit des Baugrundes bei Einwirkung vertikal gezogener Ankerplatten als zweidimensionales Bruchproblem*, 1969

12 Erbel, Klaus: *Ein Beitrag zur Untersuchung der Metamorphose von Mittelgebirgsschneedecken unter besonderer Berücksichtigung eines Verfahrens zur Bestimmung der thermischen Schneequalität*, 1969

13 Westhaus, Karl-Heinz: *Der Strukturwandel in der Binnenschiffahrt und sein Einfluß auf den Ausbau der Binnenschiffskanäle*, 1969

14 Mayer-Vorfelder, Hans-Jörg: *Ein Beitrag zur Berechnung des Erdwiderstandes unter Ansatz der logarithmischen Spirale als Gleitflächenfunktion*, 1970

15 Schulz, Manfred: *Berechnung des räumlichen Erddruckes auf die Wandung kreiszylindrischer Körper*, 1970

16 Mobasseri, Manoutschehr: *Die Rippenstützmauer. Konstruktion und Grenzen ihrer Standsicherheit*, 1970

17 Benk, Dieter: *Ein Beitrag zum Betrieb und zur Bemessung von Hochwasserrückhaltebecken*, 1970

18 Gàl, Attila: *Bestimmung der mitschwingenden Wassermasse bei überströmten Fischbauchklappen mit kreiszylindrischem Staublech*, 1971, vergriffen

19 Kuz, Klaus Dieter: *Ein Beitrag zur Frage des Einsetzens von Kavitationserscheinungen in einer Düsenströmung bei Berücksichtigung der im Wasser gelösten Gase*, 1971, vergriffen

20 Schaak, Hartmut: *Verteilleitungen von Wasserkraftanlagen*, 1971

21 *Sonderheft zur Eröffnung der neuen Versuchsanstalt des Instituts für Wasserbau der Universität Stuttgart mit Beiträgen von* Brombach, Hansjörg; Dirksen, Wolfram; Gàl, Attila; Gerlach, Reinhard; Giesecke, Jürgen; Holthoff, Franz-Josef; Kuz, Klaus Dieter; Marotz, Günter; Minor, Hans-Erwin; Petrikat, Kurt; Röhnisch, Arthur; Rueff, Helge; Schwarz, Helmut; Vollmer, Ernst; Wildenhahn, Eberhard; 1972

22 Wang, Chung-su: *Ein Beitrag zur Berechnung der Schwingungen an Kegelstrahlschiebern*, 1972

23 Mayer-Vorfelder, Hans-Jörg: *Erdwiderstandsbeiwerte nach dem Ohde-Variationsverfahren*, 1972

24 Minor, Hans-Erwin: *Beitrag zur Bestimmung der Schwingungsanfachungsfunktionen überströmter Stauklappen*, 1972, vergriffen

25 Brombach, Hansjörg: *Untersuchung strömungsmechanischer Elemente (Fluidik) und die Möglichkeit der Anwendung von Wirbelkammerelementen im Wasserbau*, 1972, vergriffen

26 Wildenhahn, Eberhard: *Beitrag zur Berechnung von Horizontalfilterbrunnen*, 1972

27 Steinlein, Helmut: *Die Eliminierung der Schwebstoffe aus Flußwasser zum Zweck der unterirdischen Wasserspeicherung, gezeigt am Beispiel der Iller*, 1972

28 Holthoff, Franz Josef: *Die Überwindung großer Hubhöhen in der Binnenschiffahrt durch Schwimmerhebewerke*, 1973

29 Röder, Karl: *Einwirkungen aus Baugrundbewegungen auf trog- und kastenförmige Konstruktionen des Wasser- und Tunnelbaues*, 1973

30 Kretschmer, Heinz: *Die Bemessung von Bogenstaumauern in Abhängigkeit von der Talform*, 1973

31 Honekamp, Hermann: *Beitrag zur Berechnung der Montage von Unterwasserpipelines*, 1973

32 Giesecke, Jürgen: *Die Wirbelkammertriode als neuartiges Steuerorgan im Wasserbau*, und Brombach, Hansjörg: *Entwicklung, Bauformen, Wirkungsweise und Steuereigenschaften von Wirbelkammerverstärkern*, 1974

33 Rueff, Helge: *Untersuchung der schwingungserregenden Kräfte an zwei hintereinander angeordneten Tiefschützen unter besonderer Berücksichtigung von Kavitation*, 1974

34 Röhnisch, Arthur: *Einpreßversuche mit Zementmörtel für Spannbeton - Vergleich der Ergebnisse von Modellversuchen mit Ausführungen in Hüllwellrohren*, 1975

35 *Sonderheft anläßlich des 65. Geburtstages von Prof. Dr.-Ing. Kurt Petrikat mit Beiträgen von*: Brombach, Hansjörg; Erbel, Klaus; Flinspach, Dieter; Fischer jr., Richard; Gàl, Attila; Gerlach, Reinhard; Giesecke, Jürgen; Haberhauer, Robert; Hafner Edzard; Hausenblas, Bernhard; Horlacher, Hans-Burkhard; Hutarew, Andreas; Knoll, Manfred; Krummet, Ralph; Marotz, Günter; Merkle, Theodor; Miller, Christoph; Minor, Hans-Erwin; Neumayer, Hans; Rao, Syamala; Rath, Paul; Rueff, Helge; Ruppert, Jürgen; Schwarz, Wolfgang; Topal-Gökceli, Mehmet; Vollmer, Ernst; Wang, Chung-su; Weber, Hans-Georg; 1975

36 Berger, Jochum: *Beitrag zur Berechnung des Spannungszustandes in rotationssymmetrisch belasteten Kugelschalen veränderlicher Wandstärke unter Gas- und Flüssigkeitsdruck durch Integration schwach singulärer Differentialgleichungen*, 1975

37 Dirksen, Wolfram: *Berechnung instationärer Abflußvorgänge in gestauten Gerinnen mittels Differenzenverfahren und die Anwendung auf Hochwasserrückhaltebecken*, 1976

38 Horlacher, Hans-Burkhard: *Berechnung instationärer Temperatur- und Wärmespannungsfelder in langen mehrschichtigen Hohlzylindern*, 1976

39 Hafner, Edzard: *Untersuchung der hydrodynamischen Kräfte auf Baukörper im Tiefwasserbereich des Meeres*, 1977, ISBN 3-921694-39-6

40 Ruppert, Jürgen: *Über den Axialwirbelkammerverstärker für den Einsatz im Wasserbau*, 1977, ISBN 3-921694-40-X

41 Hutarew, Andreas: *Beitrag zur Beeinflußbarkeit des Sauerstoffgehalts in Fließgewässern an Abstürzen und Wehren*, 1977, ISBN 3-921694-41-8, *vergriffen*

42 Miller, Christoph: *Ein Beitrag zur Bestimmung der schwingungserregenden Kräfte an unterströmten Wehren*, 1977, ISBN 3-921694-42-6

43 Schwarz, Wolfgang: *Druckstoßberechnung unter Berücksichtigung der Radial- und Längsverschiebungen der Rohrwandung*, 1978, ISBN 3-921694-43-4

44 Kinzelbach, Wolfgang: *Numerische Untersuchungen über den optimalen Einsatz variabler Kühlsysteme einer Kraftwerkskette am Beispiel Oberrhein*, 1978, ISBN 3-921694-44-2

45 Barczewski, Baldur: *Neue Meßmethoden für Wasser-Luftgemische und deren Anwendung auf zweiphasige Auftriebsstrahlen*, 1979, ISBN 3-921694-45-0

46 Neumayer, Hans: *Untersuchung der Strömungsvorgänge in radialen Wirbelkammerverstärkern*, 1979, ISBN 3-921694-46-9

47 Elalfy, Youssef-Elhassan: *Untersuchung der Strömungsvorgänge in Wirbelkammerdioden und -drosseln*, 1979, ISBN 3-921694-47-7

48 Brombach, Hansjörg: *Automatisierung der Bewirtschaftung von Wasserspeichern*, 1981, ISBN 3-921694-48-5

49 Geldner, Peter: *Deterministische und stochastische Methoden zur Bestimmung der Selbstdichtung von Gewässern*, 1981, ISBN 3-921694-49-3, *vergriffen*

50 Mehlhorn, Hans: *Temperaturveränderungen im Grundwasser durch Brauchwassereinleitungen*, 1982, ISBN 3-921694-50-7, *vergriffen*

51 Hafner, Edzard: *Rohrleitungen und Behälter im Meer*, 1983, ISBN 3-921694-51-5

52 Rinnert, Bernd: *Hydrodynamische Dispersion in porösen Medien: Einfluß von Dichteunterschieden auf die Vertikalvermischung in horizontaler Strömung*, 1983, ISBN 3-921694-52-3, *vergriffen*

53 Lindner, Wulf: *Steuerung von Grundwasserentnahmen unter Einhaltung ökologischer Kriterien*, 1983, ISBN 3-921694-53-1, *vergriffen*

54 Herr, Michael; Herzer, Jörg; Kinzelbach, Wolfgang; Kobus, Helmut; Rinnert, Bernd: *Methoden zur rechnerischen Erfassung und hydraulischen Sanierung von Grundwasserkontaminationen*, 1983, ISBN 3-921694-54-X

55 Schmitt, Paul: *Wege zur Automatisierung der Niederschlagsermittlung*, 1984, ISBN 3-921694-55-8, *vergriffen*

56 Müller, Peter: *Transport und selektive Sedimentation von Schwebstoffen bei gestautem Abfluß*, 1985, ISBN 3-921694-56-6

57 El-Qawasmeh, Fuad: *Möglichkeiten und Grenzen der Tropfbewässerung unter besonderer Berücksichtigung der Verstopfungsanfälligkeit der Tropfelemente*, 1985, ISBN 3-921694-57-4, *vergriffen*

58 Kirchenbaur, Klaus: *Mikroprozessorgesteuerte Erfassung instationärer Druckfelder am Beispiel seegangsbelasteter Baukörper*, 1985, ISBN 3-921694-58-2

59 Kobus, Helmut (Hrsg.): *Modellierung des großräumigen Wärme- und Schadstofftransports im Grundwasser*, Tätigkeitsbericht 1984/85 (DFG-Forschergruppe an den Universitäten Hohenheim, Karlsruhe und Stuttgart), 1985, ISBN 3-921694-59-0, *vergriffen*

60 Spitz, Karlheinz: *Dispersion in porösen Medien: Einfluß von Inhomogenitäten und Dichteunterschieden*, 1985, ISBN 3-921694-60-4, *vergriffen*

61 Kobus, Helmut: *An Introduction to Air-Water Flows in Hydraulics*, 1985, ISBN 3-921694-61-2

62 Kaleris, Vassilios: *Erfassung des Austausches von Oberflächen- und Grundwasser in horizontalebenen Grundwassermodellen*, 1986, ISBN 3-921694-62-0

63 Herr, Michael: *Grundlagen der hydraulischen Sanierung verunreinigter Porengrundwasserleiter*, 1987, ISBN 3-921694-63-9

64 Marx, Walter: *Berechnung von Temperatur und Spannung in Massenbeton infolge Hydratation*, 1987, ISBN 3-921694-64-7

65 Koschitzky, Hans-Peter: *Dimensionierungskonzept für Sohlbelüfter in Schußrinnen zur Vermeidung von Kavitationsschäden*, 1987, ISBN 3-921694-65-5

66 Kobus, Helmut (Hrsg.): *Modellierung des großräumigen Wärme- und Schadstofftransports im Grundwasser*, Tätigkeitsbericht 1986/87 (DFG-Forschergruppe an den Universitäten Hohenheim, Karlsruhe und Stuttgart) 1987, ISBN 3-921694-66-3

67 Söll, Thomas: *Berechnungsverfahren zur Abschätzung anthropogener Temperaturanomalien im Grundwasser*, 1988, ISBN 3-921694-67-1

68 Dittrich, Andreas; Westrich, Bernd: *Bodenseeufererosion, Bestandsaufnahme und Bewertung*, 1988, ISBN 3-921694-68-X, *vergriffen*

69 Huwe, Bernd; van der Ploeg, Rienk R.: *Modelle zur Simulation des Stickstoffhaushaltes von Standorten mit unterschiedlicher landwirtschaftlicher Nutzung*, 1988, ISBN 3-921694-69-8, *vergriffen*

70 Stephan, Karl: *Integration elliptischer Funktionen*, 1988, ISBN 3-921694-70-1

71 Kobus, Helmut; Zilliox, Lothaire (Hrsg.): *Nitratbelastung des Grundwassers, Auswirkungen der Landwirtschaft auf die Grundwasser- und Rohwasserbeschaffenheit und Maßnahmen zum Schutz des Grundwassers*. Vorträge des deutsch-französischen Kolloquiums am 6. Oktober 1988, Universitäten Stuttgart und Louis Pasteur Strasbourg (Vorträge in deutsch oder französisch, Kurzfassungen zweisprachig), 1988, ISBN 3-921694-71-X

72 Soyeaux, Renald: *Unterströmung von Stauanlagen auf klüftigem Untergrund unter Berücksichtigung laminarer und turbulenter Fließzustände*,1991, ISBN 3-921694-72-8

73 Kohane, Roberto: *Berechnungsmethoden für Hochwasserabfluß in Fließgewässern mit überströmten Vorländern*, 1991, ISBN 3-921694-73-6

74 Hassinger, Reinhard: *Beitrag zur Hydraulik und Bemessung von Blocksteinrampen in flexibler Bauweise*, 1991, ISBN 3-921694-74-4, *vergriffen*

75 Schäfer, Gerhard: *Einfluß von Schichtenstrukturen und lokalen Einlagerungen auf die Längsdispersion in Porengrundwasserleitern*, 1991, ISBN 3-921694-75-2

76 Giesecke, Jürgen: *Vorträge, Wasserwirtschaft in stark besiedelten Regionen; Umweltforschung mit Schwerpunkt Wasserwirtschaft*, 1991, ISBN 3-921694-76-0

77 Huwe, Bernd: *Deterministische und stochastische Ansätze zur Modellierung des Stickstoffhaushalts landwirtschaftlich genutzter Flächen auf unterschiedlichem Skalenniveau*, 1992, ISBN 3-921694-77-9, *vergriffen*

78 Rommel, Michael: *Verwendung von Kluftdaten zur realitätsnahen Generierung von Kluftnetzen mit anschließender laminar-turbulenter Strömungsberechnung*, 1993, ISBN 3-92 1694-78-7

79 Marschall, Paul: *Die Ermittlung lokaler Stofffrachten im Grundwasser mit Hilfe von Einbohrloch-Meßverfahren*, 1993, ISBN 3-921694-79-5, *vergriffen*

80 Ptak, Thomas: *Stofftransport in heterogenen Porenaquiferen: Felduntersuchungen und stochastische Modellierung*, 1993, ISBN 3-921694-80-9, *vergriffen*

81 Haakh, Frieder: *Transientes Strömungsverhalten in Wirbelkammern,* 1993, ISBN 3-921694-81-7

82 Kobus, Helmut; Cirpka, Olaf; Barczewski, Baldur; Koschitzky, Hans-Peter: *Versuchseinrichtung zur Grundwasser- und Altlastensanierung VEGAS, Konzeption und Programmrahmen*, 1993, ISBN 3-921694-82-5

83 Zang, Weidong: *Optimaler Echtzeit-Betrieb eines Speichers mit aktueller Abflußregenerierung*, 1994, ISBN 3-921694-83-3, *vergriffen*

84 Franke, Hans-Jörg: *Stochastische Modellierung eines flächenhaften Stoffeintrages und Transports in Grundwasser am Beispiel der Pflanzenschutzmittelproblematik*, 1995, ISBN 3-921694-84-1

85 Lang, Ulrich: *Simulation regionaler Strömungs- und Transportvorgänge in Karstaquiferen mit Hilfe des Doppelkontinuum-Ansatzes: Methodenentwicklung und Parameteridentifikation*, 1995, ISBN 3-921694-85-X, *vergriffen*

86 Helmig, Rainer: *Einführung in die Numerischen Methoden der Hydromechanik*, 1996, ISBN 3-921694-86-8, *vergriffen*

87 Cirpka, Olaf: *CONTRACT: A Numerical Tool for Contaminant Transport and Chemical Transformations - Theory and Program Documentation* -, 1996, ISBN 3-921694-87-6

88 Haberlandt, Uwe: *Stochastische Synthese und Regionalisierung des Niederschlages für Schmutzfrachtberechnungen*, 1996, ISBN 3-921694-88-4

89 Croisé, Jean: *Extraktion von flüchtigen Chemikalien aus natürlichen Lockergesteinen mittels erzwungener Luftströmung*, 1996, ISBN 3-921694-89-2, *vergriffen*

90 Jorde, Klaus: *Ökologisch begründete, dynamische Mindestwasserregelungen bei Ausleitungskraftwerken*, 1997, ISBN 3-921694-90-6, *vergriffen*

91 Helmig, Rainer: *Gekoppelte Strömungs- und Transportprozesse im Untergrund - Ein Beitrag zur Hydrosystemmodellierung*-, 1998, ISBN 3-921694-91-4, *vergriffen*

92 Emmert, Martin: *Numerische Modellierung nichtisothermer Gas-Wasser Systeme in porösen Medien,* 1997, ISBN 3-921694-92-2

93 Kern, Ulrich: *Transport von Schweb- und Schadstoffen in staugeregelten Fließgewässern am Beispiel des Neckars*, 1997, ISBN 3-921694-93-0, *vergriffen*

94 Förster, Georg: *Druckstoßdämpfung durch große Luftblasen in Hochpunkten von Rohrleitungen 1997*, ISBN 3-921694-94-9

95 Cirpka, Olaf: *Numerische Methoden zur Simulation des reaktiven Mehrkomponententransports im Grundwasser*, 1997, ISBN 3-921694-95-7, *vergriffen*

96 Färber, Arne: *Wärmetransport in der ungesättigten Bodenzone: Entwicklung einer thermischen In-situ-Sanierungstechnologie*, 1997, ISBN 3-921694-96-5

97 Betz, Christoph: *Wasserdampfdestillation von Schadstoffen im porösen Medium: Entwicklung einer thermischen In-situ-Sanierungstechnologie*, 1998, SBN 3-921694-97-3

98 Xu, Yichun: *Numerical Modeling of Suspended Sediment Transport in Rivers*, 1998, ISBN 3-921694-98-1, *vergriffen*

99 Wüst, Wolfgang: *Geochemische Untersuchungen zur Sanierung CKW-kontaminierter Aquifere mit Fe(0)-Reaktionswänden,* 2000, ISBN 3-933761-02-2

100 Sheta, Hussam: *Simulation von Mehrphasenvorgängen in porösen Medien unter Einbeziehung von Hysterese-Effekten*, 2000, ISBN 3-933761-03-4

101 Ayros, Edwin: *Regionalisierung extremer Abflüsse auf der Grundlage statistischer Verfahren*, 2000, ISBN 3-933761-04-2, *vergriffen*

102 Huber, Ralf: *Compositional Multiphase Flow and Transport in Heterogeneous Porous Media*, 2000, ISBN 3-933761-05-0

103 Braun, Christopherus: *Ein Upscaling-Verfahren für Mehrphasenströmungen in porösen Medien*, 2000, ISBN 3-933761-06-9

104 Hofmann, Bernd: *Entwicklung eines rechnergestützten Managementsystems zur Beurteilung von Grundwasserschadensfällen*, 2000, ISBN 3-933761-07-7

105 Class, Holger: *Theorie und numerische Modellierung nichtisothermer Mehrphasenprozesse in NAPL-kontaminierten porösen Medien,* 2001, ISBN 3-933761-08-5

106 Schmidt, Reinhard: *Wasserdampf- und Heißluftinjektion zur thermischen Sanierung kontaminierter Standorte*, 2001, ISBN 3-933761-09-3

107 Josef, Reinhold: *Schadstoffextraktion mit hydraulischen Sanierungsverfahren unter Anwendung von grenzflächenaktiven Stoffen*, 2001, ISBN 3-933761-10-7

108 Schneider, Matthias: *Habitat- und Abflussmodellierung für Fließgewässer mit unscharfen Berechnungsansätzen*, 2001, ISBN 3-933761-11-5

109 Rathgeb, Andreas: *Hydrodynamische Bemessungsgrundlagen für Lockerdeckwerke an überströmbaren Erddämmen*, 2001, ISBN 3-933761-12-3

110 Lang, Stefan: *Parallele numerische Simulation instätionärer Probleme mit adaptiven Methoden auf unstrukturierten Gittern*, 2001, ISBN 3-933761-13-1

111 Appt, Jochen; Stumpp Simone: *Die Bodensee-Messkampagne 2001, IWS/CWR Lake Constance Measurement Program 2001*, 2002, ISBN 3-933761-14-X

112 Heimerl, Stephan: *Systematische Beurteilung von Wasserkraftprojekten,* 2002, ISBN 3-933761-15-8, *vergriffen*

113 Iqbal, Amin: *On the Management and Salinity Control of Drip Irrigation,* 2002, ISBN 3-933761-16-6

114 Silberhorn-Hemminger, Annette: *Modellierung von Kluftaquifersystemen: Geostatistische Analyse und deterministisch-stochastische Kluftgenerierung,* 2002, ISBN 3-933761-17-4

115 Winkler, Angela: *Prozesse des Wärme- und Stofftransports bei der In-situ-Sanierung mit festen Wärmequellen,* 2003, ISBN 3-933761-18-2

116 Marx, Walter: *Wasserkraft, Bewässerung, Umwelt - Planungs- und Bewertungsschwerpunkte der Wasserbewirtschaftung,* 2003, ISBN 3-933761-19-0

117 Hinkelmann, Reinhard: *Efficient Numerical Methods and Information-Processing Techniques in Environment Water,* 2003, ISBN 3-933761-20-4

118 Samaniego-Eguiguren, Luis Eduardo: *Hydrological Consequences of Land Use / Land Cover and Climatic Changes in Mesoscale Catchments*, 2003, ISBN 3-933761-21-2

119 Neunhäuserer, Lina: *Diskretisierungsansätze zur Modellierung von Strömungs- und Transportprozessen in geklüftet-porösen Medien*, 2003, ISBN 3-933761-22-0

120 Paul, Maren: *Simulation of Two-Phase Flow in Heterogeneous Poros Media with Adaptive Methods*, 2003, ISBN 3-933761-23-9

121 Ehret, Uwe: *Rainfall and Flood Nowcasting in Small Catchments using Weather Radar*, 2003, ISBN 3-933761-24-7

122 Haag, Ingo: *Der Sauerstoffhaushalt staugeregelter Flüsse am Beispiel des Neckars - Analysen, Experimente, Simulationen -,* 2003, ISBN 3-933761-25-5

123 Appt, Jochen: *Analysis of Basin-Scale Internal Waves in Upper Lake Constance,* 2003, ISBN 3-933761-26-3

124 Hrsg.: Schrenk, Volker; Batereau, Katrin; Barczewski, Baldur; Weber, Karolin und Koschitzky, Hans-Peter: *Symposium Ressource Fläche und VEGAS - Statuskolloquium 2003, 30. September und 1. Oktober 2003,* 2003, ISBN 3-933761-27-1

125 Omar Khalil Ouda: *Optimisation of Agricultural Water Use: A Decision Support System for the Gaza Strip,* 2003, ISBN 3-933761-28-0

126 Batereau, Katrin: *Sensorbasierte Bodenluftmessung zur Vor-Ort-Erkundung von Schadensherden im Untergrund*, 2004, ISBN 3-933761-29-8

127 Witt, Oliver: *Erosionsstabilität von Gewässersedimenten mit Auswirkung auf den Stofftransport bei Hochwasser am Beispiel ausgewählter Stauhaltungen des Oberrheins*, 2004, ISBN 3-933761-30-1

128 Jakobs, Hartmut: *Simulation nicht-isothermer Gas-Wasser-Prozesse in komplexen Kluft-Matrix-Systemen*, 2004, ISBN 3-933761-31-X

129 Li, Chen-Chien: *Deterministisch-stochastisches Berechnungskonzept zur Beurteilung der Auswirkungen erosiver Hochwasserereignisse in Flussstauhaltungen,* 2004, ISBN 3-933761-32-8

130 Reichenberger, Volker; Helmig, Rainer; Jakobs, Hartmut; Bastian, Peter; Niessner, Jennifer: *Complex Gas-Water Processes in Discrete Fracture-Matrix Systems: Up-scaling, Mass-Conservative Discretization and Efficient Multilevel Solution,* 2004, ISBN 3-933761-33-6

131 Hrsg.: Barczewski, Baldur; Koschitzky, Hans-Peter; Weber, Karolin; Wege, Ralf: *VEGAS - Statuskolloquium 2004,* Tagungsband zur Veranstaltung am 05. Oktober 2004 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2004, ISBN 3-933761-34-4

132 Asie, Kemal Jabir: *Finite Volume Models for Multiphase Multicomponent Flow through Porous Media*. 2005, ISBN 3-933761-35-2

133 Jacoub, George: *Development of a 2-D Numerical Module for Particulate Contaminant Transport in Flood Retention Reservoirs and Impounded Rivers,* 2004, ISBN 3-933761-36-0

134 Nowak, Wolfgang: *Geostatistical Methods for the Identification of Flow and Transport Parameters in the Subsurface,* 2005, ISBN 3-933761-37-9

135 Süß, Mia: *Analysis of the influence of structures and boundaries on flow and transport processes in fractured porous media*, 2005, ISBN 3-933761-38-7

136 Jose, Surabhin Chackiath: *Experimental Investigations on Longitudinal Dispersive Mixing in Heterogeneous Aquifers,* 2005, ISBN: 3-933761-39-5

137 Filiz, Fulya: *Linking Large-Scale Meteorological Conditions to Floods in Mesoscale Catchments,* 2005, ISBN 3-933761-40-9

138 Qin, Minghao: *Wirklichkeitsnahe und recheneffiziente Ermittlung von Temperatur und Spannungen bei großen RCC-Staumauern*, 2005, ISBN 3-933761-41-7

139 Kobayashi, Kenichiro: *Optimization Methods for Multiphase Systems in the Subsurface - Application to Methane Migration in Coal Mining Areas*, 2005, ISBN 3-933761-42-5

140 Rahman, Md. Arifur: *Experimental Investigations on Transverse Dispersive Mixing in Heterogeneous Porous Media,* 2005, ISBN 3-933761-43-3

141 Schrenk, Volker: *Ökobilanzen zur Bewertung von Altlastensanierungsmaßnahmen,* 2005, ISBN 3-933761-44-1

142 Hundecha, Hirpa Yeshewatesfa: *Regionalization of Parameters of a Conceptual Rainfall-Runoff Model*, 2005, ISBN: 3-933761-45-X

143 Wege, Ralf: *Untersuchungs- und Überwachungsmethoden für die Beurteilung natürlicher Selbstreinigungsprozesse im Grundwasser,* 2005, ISBN 3-933761-46-8

144 Breiting, Thomas: *Techniken und Methoden der Hydroinformatik - Modellierung von komplexen Hydrosystemen im Untergrund*, 2006, ISBN 3-933761-47-6

145 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Müller, Martin: *Ressource Untergrund: 10 Jahre VEGAS: Forschung und Technologieentwicklung zum Schutz von Grundwasser und Boden,* Tagungsband zur Veranstaltung am 28. und 29. September 2005 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2005, ISBN 3-933761-48-4

146 Rojanschi, Vlad: *Abflusskonzentration in mesoskaligen Einzugsgebieten unter Berücksichtigung des Sickerraumes,* 2006, ISBN 3-933761-49-2

147 Winkler, Nina Simone: *Optimierung der Steuerung von Hochwasserrückhaltebeckensystemen,* 2006, ISBN 3-933761-50-6

148 Wolf, Jens: *Räumlich differenzierte Modellierung der Grundwasserströmung alluvialer Aquifere für mesoskalige Einzugsgebiete*, 2006, ISBN: 3-933761-51-4

149 Kohler, Beate: *Externe Effekte der Laufwasserkraftnutzung*, 2006, ISBN 3-933761-52-2

150 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Stuhrmann, Matthias: *VEGAS-Statuskolloquium 2006,* Tagungsband zur Veranstaltung am 28. September 2006 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2006, ISBN 3-933761-53-0

151 Niessner, Jennifer: *Multi-Scale Modeling of Multi-Phase - Multi-Component Processes in Heterogeneous Porous Media*, 2006, ISBN 3-933761-54-9

152 Fischer, Markus: *Beanspruchung eingeerdeter Rohrleitungen infolge Austrocknung bindiger Böden,* 2006, ISBN 3-933761-55-7

153 Schneck, Alexander: *Optimierung der Grundwasserbewirtschaftung unter Berücksichtigung der Belange der Wasserversorgung, der Landwirtschaft und des Naturschutzes*, 2006, ISBN 3-933761-56-5

154 Das, Tapash: *The Impact of Spatial Variability of Precipitation on the Predictive Uncertainty of Hydrological Models,* 2006, ISBN 3-33761-57-3

155 Bielinski, Andreas: *Numerical Simulation of CO_2 sequestration in geological formations*, 2007, ISBN 3-933761-58-1

156 Mödinger, Jens: *Entwicklung eines Bewertungs- und Entscheidungsunterstützungssystems für eine nachhaltige regionale Grundwasserbewirtschaftung*, 2006, ISBN 3-933761-60-3

157 Manthey, Sabine: *Two-phase flow processes with dynamic effects in porous media - parameter estimation and simulation,* 2007, ISBN 3-933761-61-1

158 Pozos Estrada, Oscar: *Investigation on the Effects of Entrained Air in Pipelines,* 2007, ISBN 3-933761-62-X

159 Ochs, Steffen Oliver: *Steam injection into saturated porous media – process analysis including experimental and numerical investigations*, 2007, ISBN 3-933761-63-8

160 Marx, Andreas: *Einsatz gekoppelter Modelle und Wetterradar zur Abschätzung von Niederschlagsintensitäten und zur Abflussvorhersage, 2007,* ISBN 3-933761-64-6

161 Hartmann, Gabriele Maria: *Investigation of Evapotranspiration Concepts in Hydrological Modelling for Climate Change Impact Assessment*, 2007, ISBN 3-933761-65-4

185 Zöllner, Frank: *Implementierung und Anwendung netzfreier Methoden im Konstruktiven Wasserbau und in der Hydromechanik*, 2009, ISBN 978-3-933761-89-7

186 Vasin, Milos: *Influence of the soil structure and property contrast on flow and transport in the unsaturated zone*, 2010, ISBN 978-3-933761-90-3

187 Li, Jing: *Application of Copulas as a New Geostatistical Tool*, 2010, ISBN 978-3-933761-91-0

188 AghaKouchak, Amir: *Simulation of Remotely Sensed Rainfall Fields Using Copulas*, 2010, ISBN 978-3-933761-92-7

189 Thapa, Pawan Kumar: *Physically-based spatially distributed rainfall runoff modelling for soil erosion estimation*, 2010, ISBN 978-3-933761-93-4

190 Wurms, Sven: *Numerische Modellierung der Sedimentationsprozesse in Retentionsanlagen zur Steuerung von Stoffströmen bei extremen Hochwasserabflussereignissen*, 2011, ISBN 978-3-933761-94-1

191 Merkel, Uwe: *Unsicherheitsanalyse hydraulischer Einwirkungen auf Hochwasserschutzdeiche und Steigerung der Leistungsfähigkeit durch adaptive Strömungsmodellierung*, 2011, ISBN 978-3-933761-95-8

192 Fritz, Jochen: *A Decoupled Model for Compositional Non-Isothermal Multiphase Flow in Porous Media and Multiphysics Approaches for Two-Phase Flow*, 2010, ISBN 978-3-933761-96-5

193 Weber, Karolin (Hrsg.): *12. Treffen junger WissenschaftlerInnen an Wasserbauinstituten*, 2010, ISBN 978-3-933761-97-2

194 Bliefernicht, Jan-Geert: *Probability Forecasts of Daily Areal Precipitation for Small River Basins,* 2011, ISBN 978-3-933761-98-9

195 Hrsg.: Koschitzky, Hans-Peter; Braun, Jürgen: *VEGAS-Kolloquium 2010 In-situ-Sanierung - Stand und Entwicklung Nano und ISCO -,* Tagungsband zur Veranstaltung am 07. Oktober 2010 an der Universität Stuttgart, Campus Stuttgart-Vaihingen*,* 2010*,* ISBN 978-3-933761-99-6

196 Gafurov, Abror: *Water Balance Modeling Using Remote Sensing Information - Focus on Central Asia,* 2010, ISBN 978-3-942036-00-9

197 Mackenberg, Sylvia: *Die Quellstärke in der Sickerwasserprognose: Möglichkeiten und Grenzen von Labor- und Freilanduntersuchungen,* 2010, ISBN 978-3-942036-01-6

198 Singh, Shailesh Kumar: *Robust Parameter Estimation in Gauged and Ungauged Basins,* 2010, ISBN 978-3-942036-02-3

199 Doğan, Mehmet Onur: *Coupling of porous media flow with pipe flow,* 2011, ISBN 978-3-942036-03-0

200 Liu, Min: *Study of Topographic Effects on Hydrological Patterns and the Implication on Hydrological Modeling and Data Interpolation,* 2011, ISBN 978-3-942036-04-7

201 Geleta, Habtamu Itefa: *Watershed Sediment Yield Modeling for Data Scarce Areas,* 2011, ISBN 978-3-942036-05-4

202 Franke, Jörg: *Einfluss der Überwachung auf die Versagenswahrscheinlichkeit von Staustufen,* 2011, ISBN 978-3-942036-06-1

203 Bakimchandra, Oinam: *Integrated Fuzzy-GIS approach for assessing regional soil erosion risks,* 2011, ISBN 978-3-942036-07-8

204 Alam, Muhammad Mahboob: *Statistical Downscaling of Extremes of Precipitation in Mesoscale Catchments from Different RCMs and Their Effects on Local Hydrology,* 2011, ISBN 978-3-942036-08-5

205 Hrsg.: Koschitzky, Hans-Peter; Braun, Jürgen: *VEGAS-Kolloquium 2011 Flache Geothermie - Perspektiven und Risiken,* Tagungsband zur Veranstaltung am 06. Oktober 2011 an der Universität Stuttgart, Campus Stuttgart-Vaihingen*,* 2011*,* ISBN 978-3-933761-09-2

206 Haslauer, Claus: *Analysis of Real-World Spatial Dependence of Subsurface Hydraulic Properties Using Copulas with a Focus on Solute Transport Behaviour,* 2011, ISBN 978-3-942036-10-8

207 Dung, Nguyen Viet: *Multi-objective automatic calibration of hydrodynamic models – development of the concept and an application in the Mekong Delta,* 2011, ISBN 978-3-942036-11-5

208 Hung, Nguyen Nghia: *Sediment dynamics in the floodplain of the Mekong Delta, Vietnam,* 2011, ISBN 978-3-942036-12-2

209 Kuhlmann, Anna: *Influence of soil structure and root water uptake on flow in the unsaturated zone,* 2012, ISBN 978-3-942036-13-9

210 Tuhtan, Jeffrey Andrew: *Including the Second Law Inequality in Aquatic Ecodynamics: A Modeling Approach for Alpine Rivers Impacted by Hydropeaking,* 2012, ISBN 978-3-942036-14-6

211 Tolossa, Habtamu: *Sediment Transport Computation Using a Data-Driven Adaptive Neuro-Fuzzy Modelling Approach,* 2012, ISBN 978-3-942036-15-3

212 Tatomir, Alexandru-Bodgan: *From Discrete to Continuum Concepts of Flow in Fractured Porous Media,* 2012, ISBN 978-3-942036-16-0

213 Erbertseder, Karin: *A Multi-Scale Model for Describing Cancer-Therapeutic Transport in the Human Lung,* 2012, ISBN 978-3-942036-17-7

214 Noack, Markus: *Modelling Approach for Interstitial Sediment Dynamics and Reproduction of Gravel Spawning Fish,* 2012, ISBN 978-3-942036-18-4

215 De Boer, Cjestmir Volkert: *Transport of Nano Sized Zero Valent Iron Colloids during Injection into the Subsurface*, 2012, ISBN 978-3-942036-19-1

216 Pfaff, Thomas: *Processing and Analysis of Weather Radar Data for Use in Hydrology*, 2013, ISBN 978-3-942036-20-7

217 Lebrenz, Hans-Henning: *Addressing the Input Uncertainty for Hydrological Modeling by a New Geostatistical Method*, 2013, ISBN 978-3-942036-21-4

218 Darcis, Melanie Yvonne: *Coupling Models of Different Complexity for the Simulation of CO_2 Storage in Deep Saline Aquifers*, 2013, ISBN 978-3-942036-22-1

219 Beck, Ferdinand: *Generation of Spatially Correlated Synthetic Rainfall Time Series in High Temporal Resolution - A Data Driven Approach*, 2013, ISBN 978-3-942036-23-8

220 Guthke, Philipp: *Non-multi-Gaussian spatial structures: Process-driven natural genesis, manifestation, modeling approaches, and influences on dependent processes*, 2013, ISBN 978-3-942036-24-5

221 Walter, Lena: *Uncertainty studies and risk assessment for CO_2 storage in geological formations*, 2013, ISBN 978-3-942036-25-2

222 Wolff, Markus: *Multi-scale modeling of two-phase flow in porous media including capillary pressure effects,* 2013, ISBN 978-3-942036-26-9

223 Mosthaf, Klaus Roland: *Modeling and analysis of coupled porous-medium and free flow with application to evaporation processes,* 2014, ISBN 978-3-942036-27-6

224 Leube, Philipp Christoph: *Methods for Physically-Based Model Reduction in Time: Analysis, Comparison of Methods and Application*, 2013, ISBN 978-3-942036-28-3

225 Rodríguez Fernández, Jhan Ignacio: *High Order Interactions among environmental variables: Diagnostics and initial steps towards modeling*, 2013, ISBN 978-3-942036-29-0

226 Eder, Maria Magdalena: *Climate Sensitivity of a Large Lake*, 2013, ISBN 978-3-942036-30-6

227 Greiner, Philipp: *Alkoholinjektion zur In-situ-Sanierung von CKW Schadensherden in Grundwasserleitern: Charakterisierung der relevanten Prozesse auf unterschiedlichen Skalen*, 2014, ISBN 978-3-942036-31-3

228 Lauser, Andreas: *Theory and Numerical Applications of Compositional Multi-Phase Flow in Porous Media*, 2014, ISBN 978-3-942036-32-0

229 Enzenhöfer, Rainer: *Risk Quantification and Management in Water Production and Supply Systems*, 2014, ISBN 978-3-942036-33-7

230 Faigle, Benjamin: *Adaptive modelling of compositional multi-phase flow with capillary pressure*, 2014, ISBN 978-3-942036-34-4

231 Oladyshkin, Sergey: *Efficient modeling of environmental systems in the face of complexity and uncertainty*, 2014, ISBN 978-3-942036-35-1

232 Sugimoto, Takayuki: *Copula based Stochastic Analysis of Discharge Time Series*, 2014, ISBN 978-3-942036-36-8

233 Koch, Jonas: *Simulation, Identification and Characterization of Contaminant Source Architectures in the Subsurface*, 2014, ISBN 978-3-942036-37-5

234 Zhang, Jin: *Investigations on Urban River Regulation and Ecological Rehabilitation Measures, Case of Shenzhen in China*, 2014, ISBN 978-3-942036-38-2

235 Siebel, Rüdiger: *Experimentelle Untersuchungen zur hydrodynamischen Belastung und Standsicherheit von Deckwerken an überströmbaren Erddämmen,* 2014, ISBN 978-3-942036-39-9

236 Baber, Katherina: *Coupling free flow and flow in porous media in biological and technical applications: From a simple to a complex interface description,* 2014, ISBN 978-3-942036-40-5

237 Nuske, Klaus Philipp: *Beyond Local Equilibrium — Relaxing local equilibrium assumptions in multiphase flow in porous media,* 2014, ISBN 978-3-942036-41-2

238 Geiges, Andreas: *Efficient concepts for optimal experimental design in nonlinear environmental systems*, 2014, ISBN 978-3-942036-42-9

239 Schwenck, Nicolas: *An XFEM-Based Model for Fluid Flow in Fractured Porous Media*, 2014, ISBN 978-3-942036-43-6

240 Chamorro Chávez, Alejandro: *Stochastic and hydrological modelling for climate change prediction in the Lima region, Peru*, 2015, ISBN 978-3-942036-44-3

241 Yulizar: *Investigation of Changes in Hydro-Meteorological Time Series Using a Depth-Based Approach*, 2015, ISBN 978-3-942036-45-0

242 Kretschmer, Nicole: *Impacts of the existing water allocation scheme on the Limarí watershed – Chile, an integrative approach*, 2015, ISBN 978-3-942036-46-7

243 Kramer, Matthias: *Luftbedarf von Freistrahlturbinen im Gegendruckbetrieb*, 2015, ISBN 978-3-942036-47-4

244 Hommel, Johannes: *Modeling biogeochemical and mass transport processes in the subsurface: Investigation of microbially induced calcite precipitation*, 2016, ISBN 978-3-942036-48-1

245 Germer, Kai: *Wasserinfiltration in die ungesättigte Zone eines makroporösen Hanges und deren Einfluss auf die Hangstabilität*, 2016, ISBN 978-3-942036-49-8

246 Hörning, Sebastian: *Process-oriented modeling of spatial random fields using copulas*, 2016, ISBN 978-3-942036-50-4

247 Jambhekar, Vishal: *Numerical modeling and analysis of evaporative salinization in a coupled free-flow porous-media system,* 2016, ISBN 978-3-942036-51-1

248 Huang, Yingchun: *Study on the spatial and temporal transferability of conceptual hydrological models*, 2016, ISBN 978-3-942036-52-8

249 Kleinknecht, Simon Matthias: *Migration and retention of a heavy NAPL vapor and remediation of the unsaturated zone*, 2016, ISBN 978-3-942036-53-5

250 Kwakye, Stephen Oppong: *Study on the effects of climate change on the hydrology of the West African sub-region*, 2016, ISBN 978-3-942036-54-2

251 Kissinger, Alexander: *Basin-Scale Site Screening and Investigation of Possible Impacts of CO_2 Storage on Subsurface Hydrosystems*, 2016, ISBN 978-3-942036-55-9

252 Müller, Thomas: *Generation of a Realistic Temporal Structure of Synthetic Precipitation Time Series for Sewer Applications*, 2017, ISBN 978-3-942036-56-6

253 Grüninger, Christoph: *Numerical Coupling of Navier-Stokes and Darcy Flow for Soil-Water Evaporation*, 2017, ISBN 978-3-942036-57-3

254 Suroso: *Asymmetric Dependence Based Spatial Copula Models: Empirical Investigations and Consequences on Precipitation Fields*, 2017, ISBN 978-3-942036-58-0

255 Müller, Thomas; Mosthaf, Tobias; Gunzenhauser, Sarah; Seidel, Jochen; Bárdossy, András: *Grundlagenbericht Niederschlags-Simulator (NiedSim3)*, 2017, ISBN 978-3-942036-59-7

256 Mosthaf, Tobias: *New Concepts for Regionalizing Temporal Distributions of Precipitation and for its Application in Spatial Rainfall Simulation,* 2017, ISBN 978-3-942036-60-3

257 Fenrich, Eva Katrin: *Entwicklung eines ökologisch-ökonomischen Vernetzungsmodells für Wasserkraftanlagen und Mehrzweckspeicher,* 2018, ISBN 978-3-942036-61-0

258 Schmidt, Holger: *Microbial stabilization of lotic fine sediments*, 2018, ISBN 978-3-942036-62-7

259 Fetzer, Thomas: *Coupled Free and Porous-Medium Flow Processes Affected by Turbulence and Roughness – Models, Concepts and Analysis*, 2018, ISBN 978-3-942036-63-4

260 Schröder, Hans Christoph: *Large-scale High Head Pico Hydropower Potential Assessment*, 2018, ISBN 978-3-942036-64-1

261 Bode, Felix: *Early-Warning Monitoring Systems for Improved Drinking Water Resource Protection*, 2018, ISBN 978-3-942036-65-8

262 Gebler, Tobias: *Statistische Auswertung von simulierten Talsperrenüberwachungsdaten zur Identifikation von Schadensprozessen an Gewichtsstaumauern*, 2018, ISBN 978-3-942036-66-5

263 Harten, Matthias von: *Analyse des Zuppinger-Wasserrades – Hydraulische Optimierungen unter Berücksichtigung ökologischer Aspekte*, 2018, ISBN 978-3-942036-67-2

264 Yan, Jieru: *Nonlinear estimation of short time precipitation using weather radar and surface observations*, 2018, ISBN 978-3-942036-68-9

265 Beck, Martin: *Conceptual approaches for the analysis of coupled hydraulic and geomechanical processes*, 2019, ISBN 978-3-942036-69-6

266 Haas, Jannik: *Optimal planning of hydropower and energy storage technologies for fully renewable power systems*, 2019, ISBN 978-3-942036-70-2

267 Schneider, Martin: *Nonlinear Finite Volume Schemes for Complex Flow Processes and Challenging Grids*, 2019, ISBN 978-3-942036-71-9

268 Most, Sebastian Christopher: *Analysis and Simulation of Anomalous Transport in Porous Media*, 2019, ISBN 978-3-942036-72-6

269 Buchta, Rocco: *Entwicklung eines Ziel- und Bewertungssystems zur Schaffung nachhaltiger naturnaher Strukturen in großen sandgeprägten Flüssen des norddeutschen Tieflandes*, 2019, ISBN 978-3-942036-73-3

270 Thom, Moritz: *Towards a Better Understanding of the Biostabilization Mechanisms of Sediment Beds*, 2019, ISBN 978-3-942036-74-0

271 Stolz, Daniel: *Die Nullspannungstemperatur in Gewichtsstaumauern unter Berücksichtigung der Festigkeitsentwicklung des Betons*, 2019, ISBN 978-3-942036-75-7

272 Rodriguez Pretelin, Abelardo: *Integrating transient flow conditions into groundwater well protection*, 2020, ISBN: 978-3-942036-76-4

273 Weishaupt, Kilian: *Model Concepts for Coupling Free Flow with Porous Medium Flow at the Pore-Network Scale: From Single-Phase Flow to Compositional Non-Isothermal Two-Phase Flow*, 2020, ISBN: 978-3-942036-77-1

274 Koch, Timo: *Mixed-dimension models for flow and transport processes in porous media with embedded tubular network systems*, 2020, ISBN: 978-3-942036-78-8

275 Gläser, Dennis: *Discrete fracture modeling of multi-phase flow and deformation in fractured poroelastic media*, 2020, ISBN: 978-3-942036-79-5

276 Seitz, Lydia: *Development of new methods to apply a multi-parameter approach – A first step towards the determination of colmation*, 2020, ISBN: 978-3-942036-80-1

277 Ebrahim Bakhshipour, Amin: *Optimizing hybrid decentralized systems for sustainable urban drainage infrastructures planning*, 2021, ISBN: 978-3-942036-81-8

278 Seitz, Gabriele: *Modeling Fixed-Bed Reactors for Thermochemical Heat Storage with the Reaction System CaO/Ca(OH)$_2$*, 2021, ISBN: 978-3-942036-82-5

279 Emmert, Simon: *Developing and Calibrating a Numerical Model for Microbially Enhanced Coal-Bed Methane Production*, 2021, ISBN: 978-3-942036-83-2

280 Heck, Katharina Klara: *Modelling and analysis of multicomponent transport at the interface between free- and porous-medium flow - influenced by radiation and roughness*, 2021, ISBN: 978-3-942036-84-9

281 Ackermann, Sina: *A multi-scale approach for drop/porous-medium interaction*, 2021, ISBN: 978-3-942036-85-6

282 Beckers, Felix: *Investigations on Functional Relationships between Cohesive Sediment Erosion and Sediment Characteristics*, 2021, ISBN: 978-3-942036-86-3

283 Schlabing, Dirk: *Generating Weather for Climate Impact Assessment on Lakes*, 2021, ISBN: 978-3-942036-87-0

284 Becker, Beatrix: *Efficient multiscale multiphysics models accounting for reversible flow at various subsurface energy storage sites*, 2021, ISBN: 978-3-942036-88-7

285 Reuschen, Sebastian: *Bayesian Inversion and Model Selection of Heterogeneities in Geostatistical Subsurface Modeling*, 2021, ISBN: 978-3-942036-89-4

286 Michalkowski, Cynthia: *Modeling water transport at the interface between porous GDL and gas distributor of a PEM fuel cell cathode*, 2022, ISBN: 978-3-942036-90-0

287 Koca, Kaan: *Advanced experimental methods for investigating flow-biofilm-sediment interactions*, 2022, ISBN: 978-3-942036-91-7

288 Modiri, Ehsan: *Clustering simultaneous occurrences of extreme floods in the Neckar catchment*, 2022, ISBN: 978-3-942036-92-4

289 Mayar, Mohammad Assem: *High-resolution spatio-temporal measurements of the colmation phenomenon under laboratory conditions*, 2022, ISBN: 978-3-942036-93-1

290 Schäfer Rodrigues Silva, Aline: *Quantifying and Visualizing Model Similarities for Multi-Model Methods*, 2022, ISBN: 978-3-942036-94-8

Die Mitteilungshefte ab der Nr. 134 (Jg. 2005) stehen als pdf-Datei über die Homepage des Instituts: www.iws.uni-stuttgart.de zur Verfügung.