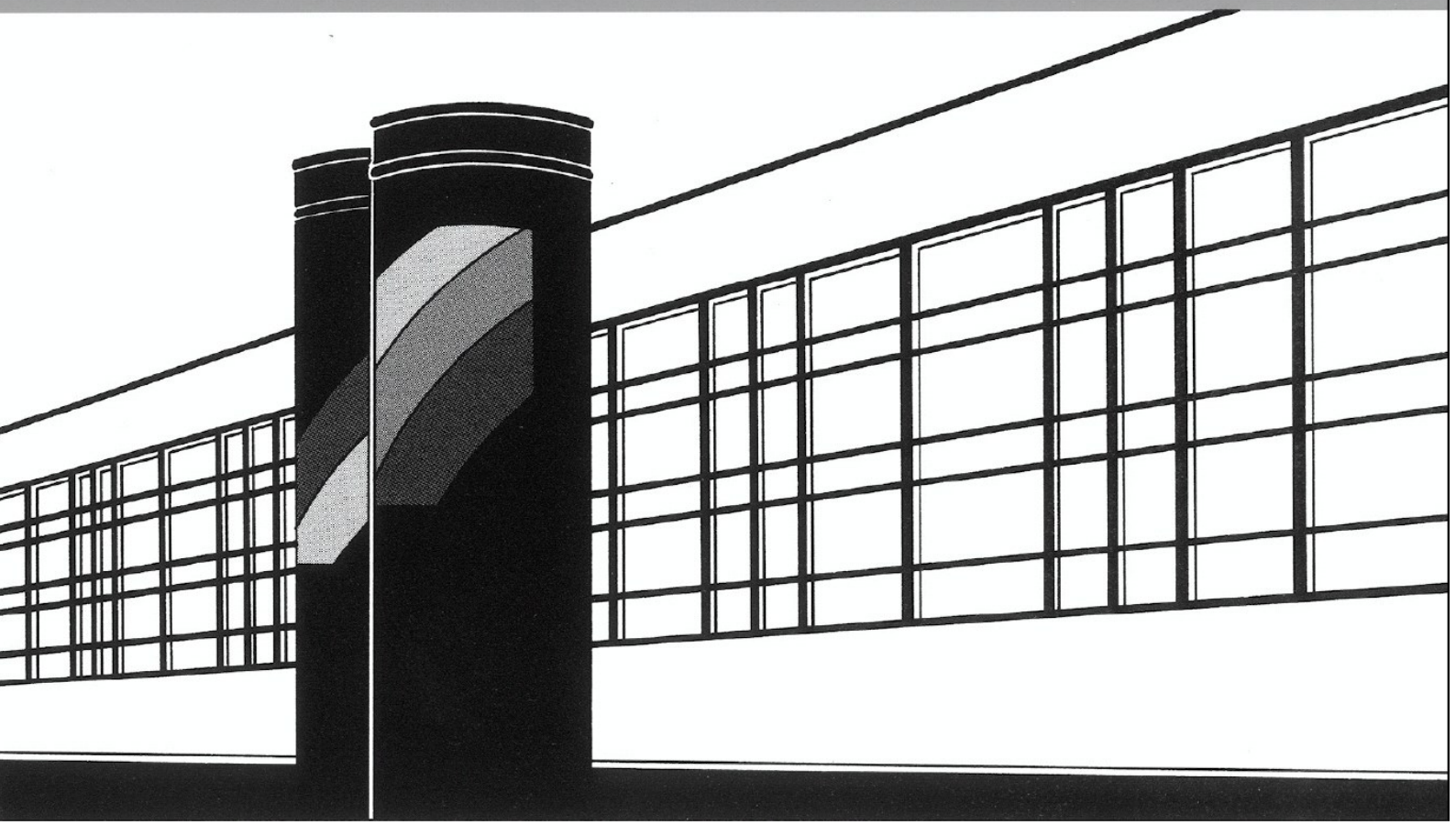
Universität Stuttgart

Institut für Wasser- und Umweltsystemmodellierung

Mitteilungen



Heft 299 Farid Mohammadi

A Surrogate-Assisted Bayesian Framework for Uncertainty-Aware Validation Benchmarks

A Surrogate-Assisted Bayesian Framework for Uncertainty-Aware Validation Benchmarks

von der Fakultät Bau- und Umweltingenieurwissenschaften
der Universität Stuttgart und dem Stuttgart Center for Simulation Science
zur Erlangung der Würde eines
Doktor-Ingenieurs (Dr.-Ing.) genehmigte Abhandlung

vorgelegt von
Farid Mohammadi
aus Rasht, Iran

Hauptberichter:	apl. Prof. Dr. rer. nat. Bernd Flemisch
Mitberichter:	apl. Prof. Dr.-Ing Sergey Oladyshkin
	Prof. Dr. Bruno Sudret

Tag der mündlichen Prüfung: 14. Februar 2023

Institut für Wasser- und Umweltsystemmodellierung
der Universität Stuttgart
2023

Heft 299

A Surrogate-Assisted Bayesian Framework for Uncertainty-Aware Validation Benchmarks

von
Dr.-Ing.
Farid Mohammadi

Eigenverlag des Instituts für Wasser- und Umweltsystemmodellierung
der Universität Stuttgart

D93 A Surrogate-Assisted Bayesian Framework for Uncertainty-Aware Validation Benchmarks

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

Mohammadi, Farid:
A Surrogate-Assisted Bayesian Framework for Uncertainty-Aware Validation Benchmarks, Universität Stuttgart. - Stuttgart: Institut für Wasser- und Umweltsystemmodellierung, 2023

(Mitteilungen Institut für Wasser- und Umweltsystemmodellierung, Universität Stuttgart: H. 299)
Zugl.: Stuttgart, Univ., Diss., 2023
ISBN 978-3-910293-03-8
NE: Institut für Wasser- und Umweltsystemmodellierung <Stuttgart>: Mitteilungen

Herausgegeben 2023 vom Eigenverlag des Instituts für Wasser- und Umweltsystem-modellierung

Druck: DCC Kästl e.K., Ostfildern

Für Mama, Oma und Onkel Hamid.

Ich widme diese Dissertation auch den iranischen Frauen, meinen kurdischen und belutschen Landsleuten.

"*Jen, Xian, Azadi*"

Frau, Leben, Freiheit!

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors, Bernd Flemisch and Sergey Oladyshkin, whose unwavering guidance, relentless support, and profound belief in my abilities have been invaluable throughout this study.

I would also like to thank the following people, without whom I would not have been able to complete this research and without whom I would not have made it through my Ph.D. study!

To my co-authors, Aline, Johannes, Martin Schneider, Elisa, Iryna, Gabi, Ilija, Wolfgang, and Holger: Thank you very much for the insights into your fields and the great cooperation.

To my student, Stefania: you have made a great contribution to the success of this work. A big thank you for that!

To my jogging pals: Felix, Ned, Sina, Katharina, Gabi, Martin Beck, Roman and Holger. It was my absolute pleasure to be able to run alongside you. I really enjoyed the discussions while we were out of breath.

I would also like to extend my sincere thanks to the team at LH^2, especially, Steffi, Beate, Pru, David and Michelle, who have been a great source of support.

To conclude, my biggest thanks to my family for all the unparalleled support they have shown me throughout my life and especially during the compilation of this dissertation.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

(a)PCE	(Arbitrary) Polynomial chaos expansion
AIC	Akaike information criterion
AIES	Affine invariant ensemble sampler
AL	Active learning
ALM	Active learning MacKay
BAL	Bayesian active learning
BaSaPCE	Bayesian sparse arbitrary polynomial chaos expansion
BHT	Bayesian hypothesis testing
BIC	Bayesian information criterion
BMA	Bayesian model averaging
BME	Bayesian model evidence
BMS	Bayesian model selection
ED	Experimental design
ED	Experimental design

EIGF	Expected improvement for global fit
KLD	Kullback–Leibler divergence
LOOCV	Leave-one-out cross validation
MAP	Maximum *a posteriori*
MCMC	Markov-Chain Monte Carlo
MH	Metropolis-Hasting algorithm
MICP	Microbially induced calcite precipitation
OLS	Ordinary least square method
OM	Original model
OMP	Orthogonal matching pursuit
PCA	Principal component analysis
PDE	Partial differential equation
PDF	Probability density function
SAED	Sequential adaptive sampling experimental design
SM	Surrogate model
SMV	Statistical model validation
SRQ	System response quantity
TOM	Theoretically optimal model
UQ	Uncertainty quantification
V&V	Verification and validation

Greek symbols

$\boldsymbol{\alpha}$	A multi-index representing the components of the multivariate polynomials
$\boldsymbol{\Sigma}$	Covariance matrix
ϵ	Discrepancy term
$\boldsymbol{\theta}$	Vector of model parameters
$\Psi(\cdot)$	Multivariate orthogonal polynomials
σ	Standard deviation
Θ	Parameter space

Number sets

$\mathbb{N}$	Natural number
$\mathbb{R}$	Real numbers

Other symbols

$\boldsymbol{c}$	Vector of PCE coefficients
$\boldsymbol{\mu}$	Mean
$\mathbb{E}[\cdot]$	Expected value
$\mathbf{y}$	Model response
$\mathcal{M}(\boldsymbol{\theta})$	Model evaluation of a parameter set
$\text{var}(\cdot)$	Variance
$\text{BF}(\cdot, \cdot)$	Bayes factor
N	Number of measurement points

N_m	Number of models
N_p	Number of model parameters
N_{ED}	Number of samples in experimental design
N_{MC}	Number of Monte-Carlo samples
$N_{D,\mathrm{spatial}}$	Number of spatial data
p	Selected polynomial degree
$P(\cdot \mid \cdot)$	Posterior probability density function
$p(\cdot \mid \cdot)$	Likelihood function
$P(\cdot)$	Prior probability density function
$P(\mathcal{M} \mid \mathcal{Y})$	Posterior model weight
$P(\mathcal{Y})$	Marginal likelihood

Roman symbols

$\mathcal{D}$	Design space
$\mathcal{M}$	Computational model
$\mathcal{N}(\cdot, \cdot)$	Normal distribution
$\mathcal{U}(\cdot, \cdot)$	Uniform distribution
$\mathcal{W}^{post}$	Expected posterior weight
$\mathcal{Y}$	Reference data set (measured or "synthetic truth")
$\tilde{\mathcal{M}}$	Surrogate model

Abstract

Over the last century, computational modeling in geoscience, especially in porous media research, has witnessed tremendous improvement. After decades of development, the state-of-the-art simulators can now solve coupled partial differential equations governing the complex subsurface multiphase flow system within a practically large spatial and temporal domain. Given the importance of computational modeling, quality assessment of these models in light of the purpose of a given simulation is of paramount importance to engineering designers and managers, public officials, and those affected by the decisions based on the predictions.

Users and developers of computational simulations deal with a challenging question: How should confidence in modeling and simulation be critically assessed? Validation is one of the primary methods for building and quantifying confidence in modeling and simulation. It investigates the degree to which a model accurately represents reality from the perspective of the intended application of the model. Usually, this comparison between model outputs and experimental data constitutes plotting the model results against data on the same axes to provide a visual assessment of agreement or lack thereof. While comparisons between model and data are at the heart of any validation procedure, there are several concerns with such naive comparisons. First, these comparisons tend to provide qualitative rather than quantitative assessments and are clearly insufficient as a basis for making decisions regarding model validity. Second, naive comparisons often disregard or only partly account for existing uncertainties in the experimental observations or the model input parameters. Third, such comparisons can not reveal whether the model is appropriate for the intended purposes, as they mainly focus on the agreement in the observable quantities.

These pitfalls give rise to the need for an uncertainty-aware framework that includes a validation metric. This metric shall provide a measure for comparison of the system

response quantities of an experiment with the ones from a computational model while accounting for uncertainties in both in a rigorous way. To address this need, we developed a statistical framework incorporating a probabilistic modeling technique using a fully Bayesian approach. The dissertation aims to help modelers perform uncertainty-aware model validation benchmarks. A two-stage Bayesian multi-model framework is discussed for modeling tasks where a set of models are at hand. To make this framework applicable for computationally demanding models, it is extended to a surrogate-assisted framework, keeping the computational costs at a reasonable level. Moreover, correction factors were introduced to compensate for the surrogate error in the Bayesian hypothesis testing and Bayesian model selection, as using surrogate representations instead of the full-fidelity computational models introduces additional errors to the validation metrics.

In this dissertation, I show how the Bayesian formalism could be materialized by employing the concept of polynomial chaos expansion to achieve more accurate surrogates with a sparse representation and account for the uncertainty in the surrogate's predictions. I also highlight how such surrogate models could be constructed with as few simulations as the computational budget allows. To this end, sequential adaptive sampling strategies are discussed, in which one attempts to augment the initial design iteratively. By doing so, informative regions in the parameter space are adequately explored. These regions are more likely to provide valuable information on the behavior of the original model responses. Using a sequential sampling strategy avoids the waste of computational resources, as opposed to the so-called one-shot designs. A series of benchmark studies are conducted to investigate the predictive capabilities of different sparsity and sequential adaptive sampling methods.

Moreover, I introduce *BayesValidRox*, an open-source, object-oriented Python package that provides an automated workflow for surrogate-based sensitivity analysis, Bayesian calibration, and validation of computational models with a modular structure. The uncertainty-aware validation framework was applied to a range of cases in the field of subsurface hydro-system modeling, mainly to flow and transport in porous media, such as flow simulation models in fractured porous media, coupling free flow and porous medium flow, and microbially induced calcite precipitation. However, this validation framework can be transferred to other disciplines in which models are used for prediction.

Zusammenfassung

Im Laufe des letzten Jahrhunderts hat die rechnergestützte Modellierung in den Geowissenschaften, insbesondere in der Erforschung poröser Medien, enorme Verbesserungen erfahren. Nach jahrzehntelanger Entwicklung sind die modernsten Simulatoren heute in der Lage, gekoppelte partielle Differentialgleichungen zu lösen, die das komplexe unterirdische Mehrphasenströmungssystem in einem großen räumlichen und zeitlichen Bereich bestimmen. In Anbetracht der Bedeutung der computergestützten Modellierung ist die Qualitätsbewertung dieser Modelle im Hinblick auf den Zweck einer bestimmten Simulation von größter Bedeutung für Ingenieure und Manager, öffentliche Entscheidungsträger und diejenigen, die von den auf den Vorhersagen basierenden Entscheidungen betroffen sind.

Nutzer und Entwickler von Computersimulationen stehen vor einer schwierigen Frage: Wie sollte das Vertrauen in Modellierung und Simulation kritisch bewertet werden? Die Validierung ist eine der wichtigsten Methoden zum Aufbau und zur Quantifizierung des Vertrauens in die Modellierung und Simulation. Es wird untersucht, inwieweit ein Modell die Realität aus der Perspektive der beabsichtigten Anwendung des Modells korrekt wiedergibt. Dieser Vergleich zwischen den Modellergebnissen und den experimentellen Daten besteht in der Regel darin, dass die Modellergebnisse den Daten auf denselben Achsen gegenübergestellt werden, um eine visuelle Bewertung der Übereinstimmung oder des Fehlens einer solchen zu ermöglichen.

Obwohl Vergleiche zwischen Modell und Daten das Herzstück eines jeden Validierungsverfahrens sind, gibt es Bedenken gegenüber solchen naiven Vergleichen. Erstens liefern diese Vergleiche eher qualitative als quantitative Bewertungen und sind als Grundlage für Entscheidungen über die Gültigkeit von Modellen eindeutig unzureichend. Zweitens lassen naive Vergleiche oft bestehende Unsicherheiten in den experimentellen Beobachtungen oder den Modellinputparametern außer Acht oder berücksichtigen sie nur

teilweise. Drittens können solche Vergleiche nicht aufzeigen, ob das Modell für die beabsichtigten Zwecke geeignet ist, da sie sich hauptsächlich auf die Übereinstimmung der beobachtbaren Größen konzentrieren.

Aus diesen Einschränkungen ergibt sich die Notwendigkeit eines Frameworks, der die Unsicherheiten berücksichtigt und eine Validierungsmetrik enthält. Diese Metrik soll ein Maß für den Vergleich der Systemreaktionsgrößen eines Experiments mit denen eines Berechnungsmodells bieten, wobei die Unsicherheiten in beiden Modellen streng zu berücksichtigen sind. Um diesen Bedarf zu decken, haben wir ein statistisches Framework entwickelt, der eine probabilistische Modellierungstechnik mit einem vollständig Bayes'schen Ansatz beinhaltet.

Diese Dissertation zielt darauf ab, Modellierer bei der Durchführung von Benchmarks zur Modellvalidierung unter Berücksichtigung von Unsicherheiten zu unterstützen. Es wird ein zweistufiges Bayes'sches Multi-Modell-Framework für Modellierungsaufgaben diskutiert, bei denen eine Reihe von Modellen zur Verfügung steht. Um dieses Framework für rechenintensive Modelle anwendbar zu machen, wird er zu einem Surrogat-gestützten Framework erweitert, der die Rechenkosten auf einem vernünftigen Niveau hält. Darüber hinaus wurden Korrekturfaktoren eingeführt, um den Surrogatfehler bei der Bayes'schen Hypothesenprüfung und der Bayes'schen Modellauswahl auszugleichen, da die Verwendung von Surrogatdarstellungen anstelle von originalgetreuen Berechnungsmodellen zusätzliche Fehler in die Validierungsmetriken einbringt.

In dieser Dissertation wird gezeigt, wie der Bayes'sche Formalismus durch den Einsatz des Konzepts der polynomialen Chaos-Expansion materialisiert werden kann, um genauere Surrogate mit einer dünnbesetzter Darstellung zu erhalten und die Unsicherheit in den Vorhersagen des Surrogats zu berücksichtigen. Es wird auch gezeigt, wie solche Ersatzmodelle mit so wenigen Simulationen wie möglich erstellt werden können. Zu diesem Zweck werden sequentielle adaptive Sampling-Strategien diskutiert, bei denen man versucht, das anfängliche Trainings-Set iterativ zu ergänzen. Auf diese Weise werden informative Regionen im Parameterraum angemessen untersucht. Diese Regionen liefern mit größerer Wahrscheinlichkeit wertvolle Informationen über das Verhalten der ursprünglichen Modellantworten. Durch die Verwendung einer sequenziellen Sampling-Strategie wird die Verschwendung von Rechenressourcen im Gegensatz zu den so genannten One-Shot-Designs vermieden. Es werden eine Reihe von Benchmark-Studien

durchgeführt, um die Vorhersagefähigkeiten verschiedener Löser für Sparse-PCE und sequentieller adaptiver Sampling-Methoden zu untersuchen.

Außerdem stelle ich *BayesValidRox* vor, ein objektorientiertes Open-Source-Python-Paket, das einen automatisierten Ablauf für Surrogat-basierte Sensitivitätsanalysen, Bayes'sche Kalibrierung und Validierung von Berechnungsmodellen mit einer modularen Struktur bietet. Das Unsicherheit-bewusste Validierung-Framework wurde auf eine Reihe von Anwendungsfällen im Bereich der Modellierung von Hydro-Systemen im Untergrund angewandt, hauptsächlich auf Strömung und Transport in porösen Medien, wie z. B. Strömungssimulationsmodelle in geklüfteten porösen Medien, Kopplung von freier Strömung und Strömung in porösen Medien sowie mikrobiell induzierte Kalzitausfällung. Dieser Validierungsrahmen kann jedoch auch auf andere Disziplinen übertragen werden, in denen Modelle zur Vorhersage verwendet werden.

1 Introduction

Modeling assists in the conceptualization and exploration of the processes' behavior and their interaction to better understand these and generate hypotheses concerning them. Moreover, it can facilitate the development of numerical experiments to test hypotheses and predict outcomes. In scientific research, a model is an abstraction of a system that describes a complex phenomenon in the most straightforward way that is adequate for the aim of modeling. This simplification contains only those components believed to be important to the problem at hand. That means a model is affected by the modelers' perception of the actual system. Since the 1950s, modeling has witnessed significant growth as a research activity. This growth is primarily due to conceptual developments in modeling techniques and advances in computational power. This advancement in modeling has also led to increased scientific understanding and identification of more complex interactions and relationships between processes and cause-effect relationships.

There has been a growing need to study environmental systems, especially subsurface systems. Subsurface environmental systems are open and complex, in which intricate hydrologic, microbiologic, mechanical, and geochemical processes occur and interact at multiple scales. Understanding and predicting system responses to natural forces (e.g., climate changes) and human activities (e.g., contaminant remediation and CO_2 sequestration) is indispensable for managing water resources, cleaning subsurface contamination and providing expert analysis to inform policy-making on long-term stewardship of nuclear waste disposal and CO_2 storage sites. Environmental scientists use modeling as an essential instrument to study the causes and effects of environmental problems, with an increasing level of detail and complexity. This tool helps engineers to approximate complex processes in nature and give sufficiently accurate estimates, predictions, and forecasts. These predictions are advantageous in decision-making, which is the basis for developing public policy, preparing safety procedures, and establishing legal liability.

1.1 Quality assessment of computational models

Computational modeling in geoscience, particularly in porous media research, has evolved tremendously over the last century. With decades of development under their belts, modern simulators are now able to solve coupled partial differential equations governing the complex subsurface multiphase flow system within a practically large spatial and temporal domain. Given the importance of computational modeling, assessment of the quality of models in light of the purpose of a given simulation is of paramount importance to engineering designers and managers, public officials, and those affected by the decisions based on the predictions. Various complementary measures exist for this quality assessment, such as unit testing of individual components of the computational model, comparisons with analytical solutions, or other computational models.

While assessing the simulation results with the measures mentioned above could hint us towards the correctness of the model, measuring the computational model performance against experimental data can enhance the confidence in the models. However, comparisons with experimental data can be highly cumbersome since the experiment in question has probably not been designed to meet the objective of the validation of a computational model. Moreover, disagreement with experimental data can have other reasons than the deficiency of the simulation code, such as an insufficient description of the experimental set-up or considerable uncertainties associated with the measurements. On the other hand, pure code intercomparison studies cannot ensure that a successfully participating model maps reality. Users and developers of computational simulations deal with a challenging question: How should confidence in modeling and simulation be critically assessed?

1.1.1 Model assessment tools

To answer the question posed above, let us look at the essential components of model development. Figure 1.1 provides a comprehensive overview of the modeling steps and the tools for model assessment.

Conceptual model is the idealized representation of the physical behavior of the reality of interest. It comprises all relevant information concerning the system of interest, modeling assumptions, simplifications, and approximations regarding the processes of

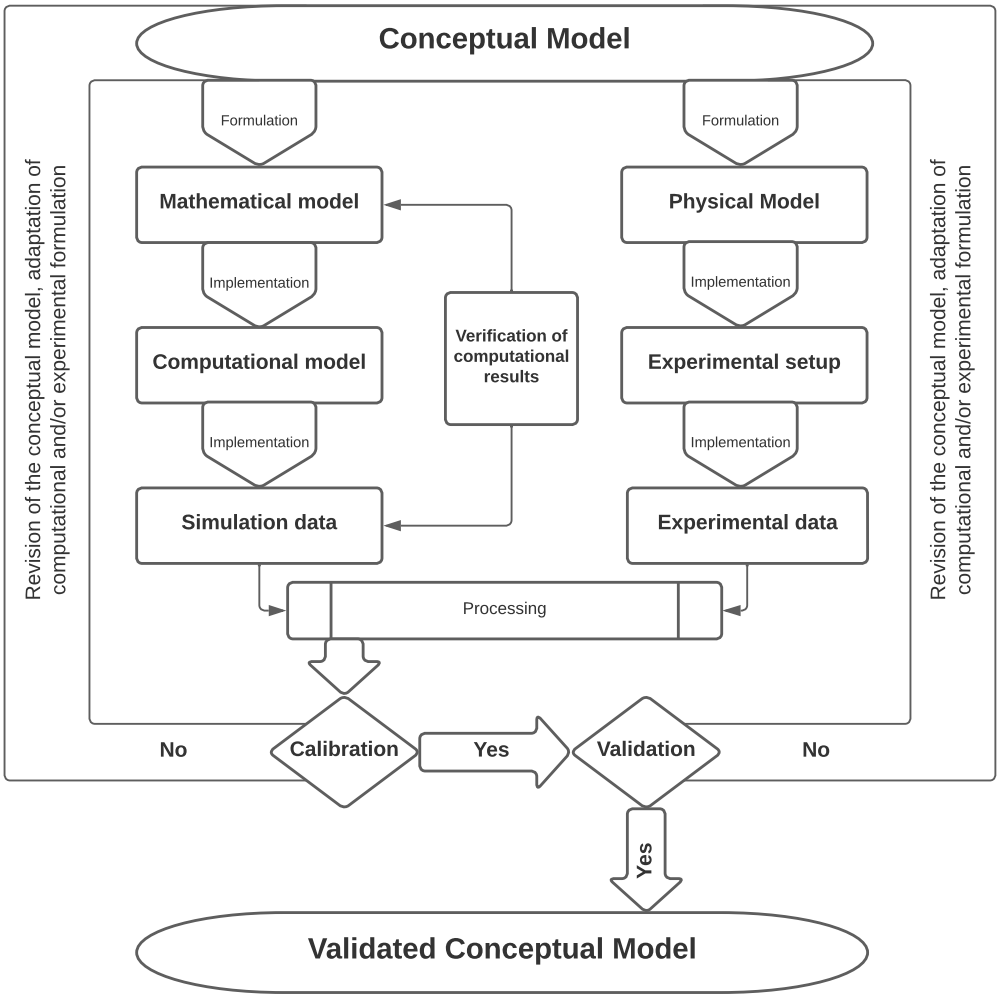


Figure 1.1: The process of model development

interest within the system, as well as the specification of the interaction of the surroundings with the system of interest. It is heavily dependent on the modeler's perception of the governing processes of the system under investigation. The fundamentals of the conceptual models are formulated in two forms: *physical model* and *mathematical model*. While the former is a smaller or larger copy of the system under investigation, the latter describes the physical processes represented in this conceptual model using a set of variables and a set of equations that establish relationships between the variables. A typical *mathematical model* contains the elements such as governing equations, supplementary sub-models (defined by constitutive equations), assumptions, and constraints (e.g., initial and boundary conditions).

A *computational model* is the numerical implementation of the mathematical model that will be solved on a computer to yield the computational predictions of the system response. This model allows scientists to conduct thousands of simulated experiments by a computer given that its execution is not time-consuming. On the other spectrum, an *experimental setup* can be established in a laboratory to resemble the *physical model*. Computational models allow you to control many variables much more precisely than real systems, and you can replicate results exactly. As a consequence, different components can be explored in ways that would otherwise be impossible in an *experimental setup*.

In what follows, the tools available to critically assess the accuracy of the models are discussed:

Verification is the process of examining whether the results of a computational model are close enough to the corresponding solutions of the underlying mathematical model. The fundamental strategy is to identify, quantify and reduce errors caused by mapping the mathematical model to a computer code. Objective verification of the mathematical model does not address the question whether it has any relationship to reality. Two types of verification are generally recognized in computational modeling: code verification and solution verification. Oberkampf et al. [2004] divided the code verification into numerical algorithm verification and software quality assurance. The primary goal of numerical algorithm verification is to accumulate sufficient evidence to demonstrate that the numerical algorithms in the code are implemented correctly and functioning as intended.

Software quality assurance deals with determining whether or not the code as a software system is reliable (implemented correctly). It provides reproducible results on specified computer hardware and a specified system with a selected software environment, including compilers, libraries, etc. Solution verification, a.k.a. numerical error estimation, deals with the quantitative assessment of the numerical accuracy of a given solution to the partial differential equations (PDEs) . Assessment of numerical accuracy is a critical issue in computations used for validation activities and in the application of the code for the intended application.

Calibration is the process of tuning parameters in the computational model to achieve some degree of agreement with experimental data. This approach has proven effective as it permits the estimation of the poorly known model compartments in the computational models. Calibration is fundamentally an optimization (estimation) problem. To specify a calibration problem, one has to define calibration parameters, targets, and objective functions.

- *Calibration parameters* are either those weakly identifiable physical attributes whose true values cannot be measured in physical experiments, or tune adjustment parameters of the computer code with no physical meaning to represent a better surrogate for the physical process.

- *Calibration targets* are the data against which the model output is compared. Calibration aims to select parameter values that produce model outputs that are "close" to the calibration targets (while "close" may be assessed graphically or visually, it is preferably encoded quantitatively by an objective function). Calibration targets are chosen based on model quantities of interest, availability of "high-quality" data, and modeling goals.

- *Objective functions* are typically scalar functions of the calibration parameters used to quantitatively assess the "closeness" of the estimated and true values. Common choices include a distance of the calibration target data from model outputs, a convexity-preserving distance transformation, or a likelihood or pseudo-likelihood. Example distances are the sum of absolute or squared differences between model outputs and calibration targets (L^1- and L^2-norms, respectively). Examples of convexity-preserving transformations of distances are various chi-squared statistics. Distances are convex objective functions; for many problems, the likelihood and pseudo-likelihood are also convex. As mentioned above, the convexity of the objective function is an essential property because convex problems are much easier to solve than non-convex ones [Dahabreh et al., 2017].

Solving the calibration problem requires optimization algorithms. These algorithms search for values of calibration parameters in the feasible domain, i.e., parameter space, that optimize the objective function. They range from simple manual tuning to more sophisticated statistical approaches such as Markov-Chain Monte Carlo (MCMC) . The latter will be discussed in Chapter 2.

Validation is an assessment of model accuracy relative to available experimental data. Its focus is on assessing the error due to the approximations and assumptions made in formulating the conceptual and mathematical models. It investigates how accurately a computational model describes reality. Validation deals with a broader range of issues:

- Assessment of the fidelity of the mathematical modeling of physical processes.
- Evaluation of the consistency, or relevance, of the mathematical model to the physical experiment being conducted.
- The influence of the experimental diagnostic techniques on the measurements themselves.
- Estimation of experimental measurement uncertainty.

Validation rests on evidence that the appropriate experiments were executed correctly and on evidence that supports the mathematical accuracy of the computed solution. These issues are practically coupled in nontrivial ways in complex validation problems, although they are logically distinct.

Comparing computational results with experimental data is a common practice in all fields of engineering and science. Researchers usually make graphical comparisons by plotting some computational system response quantity (SRQ) with the experimentally measured response over a range of input parameters. SRQs can ideally range from globally integrated to local quantities. Generally, the computational model is considered 'validated' if the computational results agree with experimental data. With this graphical comparison, the effects of uncertainty in the computational results and experimental data are not typically quoted, nor is its statistical character quantified. To address these issues, methods for quantitative comparison, i.e., validation metrics, have gained interest in the research community. These validation metrics use statistical procedures to compare the simulation results with the experimental data quantitatively. They include uncertainties and variability due to the range of the independent variables, such as a spatial coordinate or time.

Verification and validation (V&V) of computational simulations are the primary methods for building and quantifying confidence in modeling and simulation [Oberkampf and Roy, 2010]. While verification comprises the process of determining that a model implementation accurately represents the developer's conceptual description of the model

and the solution to the model, validation investigates the degree to which a model is an accurate representation of reality from the perspective of the intended application of the model [Oberkampf and Roy, 2010]. As Roache [1998] succinctly states, "Verification deals with mathematics; validation deals with physics." Model calibration can be regarded as a valuable and pragmatic path forward for applying the calibrated model in future predictions similar to the experimental database. However, calibration rarely addresses the underlying weaknesses of the models. These weaknesses are typically attributed to many modeling approximations, or deficiencies, that could be contributing to the disagreement [Oberkampf and Trucano, 2002]. The calibration process must not be undertaken as a replacement but as a response to V&V assessment [Trucano et al., 2006]. Refsgaard and Henriksen [2004] state that "*Validation tests against independent data that have not also been used for calibration are necessary in order to be able to document the predictive capability of a model.*" In other words, the major validation challenge is assessing the model in a "blind" test with experimental data. In contrast, the critical calibration issue is adjusting the physical modeling parameters to improve agreement with experimental data.

1.2 Challenges in model assessment

In scientific modeling, it is relatively common to use the word validation to refer to simple comparisons between model outputs and experimental data. Usually, this comparison constitutes plotting the model results against data on the same axes to provide a visual assessment of agreement or lack thereof. According to Moser and Oliver [2015], while comparisons between model and data are at the heart of any validation procedure, there are several concerns with such naive comparisons. First, these comparisons tend to provide qualitative rather than quantitative assessments. Such qualitative assessments are often essential and informative. However, they are insufficient for making decisions regarding model validity. Second, naive comparisons often disregard or only partly account for existing uncertainties in the experimental observations or the model input parameters. It is impossible to appropriately determine whether the model and data agree without accounting for these uncertainties. Third, such comparisons can not reveal whether the model is appropriate for the intended purposes, as they mainly focus on the agreement in the observable quantities.

These pitfalls of straightforward but naive comparisons give rise to the need for a framework that includes a validation metric. This metric must compare the system response quantities of an experiment with the ones from a computational model while accounting for uncertainties in both in a rigorous way. Probability theory methods are widely acknowledged and well-developed approaches to handle uncertainties in model validation. However, these methods require statistical convergence to produce reliable results. This convergence can be achieved when many simulations are run for different settings. Computationally expensive models with a runtime of minutes to days render the probability-based methods intractable. Therefore, the computational cost of the model validation also needs to be accounted for, especially when an uncertainty-aware model validation is envisaged.

The rest of this section is dedicated to addressing the challenges introduced above. First, a discussion on the difference between errors and uncertainties is provided, followed by all uncertainty forms. Second, validation metrics, their properties, and existing approaches are introduced. The last part of this subsection deals with the computational challenge of using probabilistic approaches for model validation.

1.2.1 Uncertainties and errors

Complex subsurface environmental systems are never entirely predictable. Model predictions of the subsurface system are inherently uncertain. This uncertainty is one of the most significant major obstacles in simulating subsurface environments. A critical component of assessing a model's predictive power is quantifying the complex system's predictive uncertainty, assisting decision-making, and guiding the experimental design and data collection to reduce uncertainty. Uncertainty quantification (UQ) is an essential part of the model assessment via V&V, both in terms of experimental measurement uncertainty and computational simulation. That is because input quantities needed from the experiment are either unavailable or imprecisely characterized.

In a computational model, uncertainties can either be random variables, such as measurement uncertainty in experiments, or unknown quantities that are not measured in experiments but are needed for input, e.g., boundary conditions. However, the main cause of errors is numerical solutions inaccuracy, such as lack of convergence of the

spatial grid and lack of time-step resolution in unsteady phenomena [Oberkampf and Barone, 2006].

Analysts commonly believe that uncertainty may manifest in aleatory and epistemic forms. The word "aleatoric" is derived from the Latin *alea* or dice, referring to a game of chance. Aleatory uncertainty arises from randomness or unpredictability due to stochasticity. It includes variability over time, across space, or among individual components in a system. It represents unknowns that lead to changes in outcomes each time we run the same experiment. Epistemic uncertainty, also known as systematic uncertainty, stems from imperfect knowledge, ignorance, and limitations to information. It includes data censoring and other forms of measurement uncertainty and often doubts about the correct form of the model itself.

Aleatoric and epistemic uncertainty can also co-occur in a single term, for example, when experimental parameters show aleatoric uncertainty and are passed to a computer simulation. A surrogate model trained using these computer simulations for uncertainty quantification will then show epistemic uncertainty dependent on or interacting with the aleatoric uncertainty of the experimental parameters [Ranftl et al., 2021]. Rather than being confined to aleatoric or epistemic uncertainty, such uncertainty is more generally inferential.

UQ intends to express both types of uncertainty separately. Quantifying the aleatoric uncertainties can be relatively straightforward, where the well-developed and widely acknowledged (frequentist) probability theory is the most basic form. Recently, techniques such as the Monte Carlo method, Karhunen–Loève, and polynomial chaos are frequently used for quantifying aleatoric uncertainties. These methods estimate an SRQ by a probability distribution, represented by its moments (in the Gaussian case, the mean and covariance suffice). Regarding the estimation of epistemic uncertainties, Bayesian probability is generally used to understand epistemic uncertainty, where probabilities indicate how certain a rational person can be about a claim. Still, there has been controversy as to whether other methods might be needed to properly or fully treat epistemic uncertainty [Oberkampf et al., 2001, Nikolaidis and Haftka, 2001, Ferson et al., 2003].

1.2.2 Quantitative model validation

According to Oberkampf and Trucano [2008], we should build predictive capability directly on quantitatively assessed model accuracy instead of making vague or ambiguous declarations that the model is "valid" or a foundation built on the calibration of the model to all available data. Oberkampf and Trucano [2000, 2002] believe that comparing computational results to experimental results should be viewed as evaluating a computable measure or a collection of these measures. The input data to the metric are the computational results and the experimental measurements of the same SRQ of interest. The authors refer to these types of measures as a *validation metric* and recommend that both uncertainties and errors be included in comparing computational and experimental results. The challenge in quantitative model assessment is how to include the uncertainties and errors in the computational model and experimental data in the calculation of validation metrics.

Different types of validation metrics have been studied in the literature to express the accuracy of a computational model through comparison of its prediction against observed data and to determine whether the model is adequate for its intended use. Coleman and Stern [1997] and Oberkampf and Trucano [2002] discuss conceptual and practical aspects of model validation and provided guidelines for conducting validation experiments and developing validation metrics. Available approaches for quantitative model validation are based on statistical confidence intervals [Oberkampf and Barone, 2006], computing distance between the model prediction and experimental data by computing the area metric [Ferson et al., 2008, Sankararaman et al., 2011], normalizing residuals [Hills and Leslie, 2003], classical statistics-based hypothesis testing [Urbina et al., 2003], Bayesian hypothesis testing [Gelfand and Dey, 1994, Zhang and Mahadevan, 2003, Mahadevan and Rebba, 2005, Geweke, 2007, Sankararaman and Mahadevan, 2011], and reliability analysis-based techniques [Rebba and Mahadevan, 2008, Sankararaman and Mahadevan, 2013, Thacker and Paez, 2014]. Liu et al. [2011], and Ling and Mahadevan [2013] investigated several of these validation approaches in detail and discussed their practical implications in engineering.

Statistical model validation (SMV) can be used to assess the prediction of computational models based on observed data in a statistical fashion [Hills and Trucano, 1999, Oberkampf and Barone, 2006, Oberkampf and Roy, 2010, Kat and Els, 2012, Sankararaman and Mahadevan, 2015, Lee et al., 2019, Kim and Youn, 2019]. SMV

requires statistical validation metrics and hypothesis testing as tools for statistical decision-making of whether a computational model is valid or not. A statistical validation metric measures the discrepancy between predicted and observed results that is used for the decision-making process [Dowding, 2001, Paez and Urbina, 2002, Hills and Trucano, 2002, Oberkampf and Trucano, 2002, Hills and Leslie, 2003, Dowding and Rutherford, 2003, Chen et al., 2004, Dowding et al., 2004, Oberkampf et al., 2004, Trucano et al., 2006, Oberkampf and Barone, 2006, Mahadevan and Rebba, 2005, Hills, 2005, Rebba et al., 2006, Xiong et al., 2009, Thonhofer et al., 2014, Zhao et al., 2017].

Scholars in related fields have discussed the features desired for statistical validation metrics according to specific scenarios [Liu et al., 2011, Ling and Mahadevan, 2013, Bi et al., 2017, Maupin et al., 2018]. These studies deal with the extended capability of statistical validation metrics for use with observation data that have particular characteristics. However, the validation metrics used in existing studies tend to be very sensitive to the distance between the observation and prediction and only weakly consider the discrepancy of the distribution with respect to the distributed degree of each dataset. For an accurate SMV, statistical validation metrics must quantify the difference between the distributions with respect to not only the difference in mean but also the variance. A discrepancy of variance between observation and prediction can occur in various situations. For example, when the uncertainty of the predicted response varies according to an input parameter, the distributions of the observed and predicted results would have a difference in variance. This dissertation focuses on statistical validation metrics, which enable us to capture the variance difference between observation and prediction to reduce decision errors in validation.

Statistical validation metrics play a crucial role in SMV; they can change the result of the decision-making process. Under various uncertainty sources, SMV can statistically adopt several previously developed validation metrics. Previous studies adopted the distribution function's shape to consider a global region discrepancy [Kullback, 1997, Hills and Trucano, 2002, Mahadevan and Rebba, 2005, Oberkampf and Barone, 2006, Ferson et al., 2008, Jeon et al., 2015]. The area metric is one popular metric. Developed by Ferson et al. [2008], the area metric generally employs a U-pooling method to solve practical problems using SMV. The likelihood is a metric for the goodness of fit, which measures the fidelity of a dataset to a designated distribution [Hills and Trucano, 2002, Oberkampf and Roy, 2010, Keysers et al., 2020]. In other work, the Bayes factor has been used to quantify the likelihood ratio of two possible distribution models for

given data [Kass and Raftery, 1995, Berger and Mortera, 1999, Keysers et al., 2020]. Kullback–Leibler divergence (KLD) has been used to imply the relative entropy of a probability density function with respect to another reference distribution [Kullback, 1997, Smith et al., 2006, Pérez-Cruz, 2008]. Probability of separation quantifies the separated degree of two distributions and is generally used for classifying two datasets [Jeon et al., 2015]. Probability residual is the squared form of the area discrepancy between two probability density functions [Oh et al., 2019, Son et al., 2020].

Hypothesis testing evaluates the plausibility of rejecting a null hypothesis according to a designated confidence level [Naylor and Finger, 1967, Balci and Sargent, 1982, Koch, 1999, Johnson et al., 2000, Wilcox, 2011, Ross, 2020] and has been employed in validation studies [Hills and Trucano, 2002, Hills and Leslie, 2003, Dowding and Rutherford, 2003, Dowding et al., 2004, Chen et al., 2004]. In this approach, the validation assessment is formulated as a decision on whether the model's predictions are consistent with the available empirical information. In SMV, the null hypothesis is that there is no significant discrepancy between the prediction and observation. Hypothesis testing constructs a distribution of the validation metric, assuming that the observation data belong to the predicted distribution. When the value of the computed validation metric exists within the confidence level of this validation metric's distribution, the hypothesis testing cannot reject the null hypothesis. Using this information, SMV determines the validity of the computational model.

Hypothesis testing approaches can be divided into two groups: classical and Bayesian hypothesis testing. Classical hypothesis testing is well established and has been explained in detail in many statistics textbooks, [e.g., Lehmann et al., 2005]. In classical hypothesis testing, a test statistic is first formulated, and the probability distributions of this statistic under the null (H_0) and alternative (H_1) hypotheses are derived theoretically or by approximations. Based on the validation data, one can compute the value of the test statistic and thus the corresponding p-value. This value is the probability that the test statistic falls outside the range defined by its computed value under the null hypothesis. In general, p-values can be considered as indicators of the quality of a null hypothesis, because a better null hypothesis corresponds to a narrower range defined by the test statistic's computed value, so there is a higher chance the test statistic will fall outside that range [Ling and Mahadevan, 2013]. Note that failing to reject the null hypothesis indicates that the accuracy of the model is acceptable, but it does not prove that the null hypothesis is true.

In the past decade, alternatively, Bayesian methods have been developed to determine the predictive capabilities of computational models (see e.g., Kennedy and O'Hagan [2001], Chen et al. [2004, 2007], Mahadevan and Rebba [2005], Rebba et al. [2006], Rebba and Mahadevan [2008], Jiang and Mahadevan [2007, 2008a,b, 2009a,b]). Bayesian hypothesis testing (BHT) incorporates the use of the Bayes factor (or factors). This factor is interpreted as the ratio of the relative likelihood of the null hypothesis that the experimental data support the model predictions divided by the alternative hypothesis that the data does not support the prediction. Bayes factors aid the decision regarding accepting or rejecting the null hypothesis test. However, as pointed out by some researchers [Ferson et al., 2008, Liu et al., 2011, Oberkampf and Barone, 2006, Oberkampf and Roy, 2010], it is not clear how a designer or policy maker should interpret this factor in a decision context concerning how much error is incurred by using the model.

One significant difference between the Bayesian and classical hypothesis testing approaches is that the Bayesian approach focuses on accepting the null hypothesis. In contrast, Edwards et al. [1963] claims that the classical approaches often tend to reject the null hypothesis based on data that do not severely distract its credibility. This is because these methods do not consider whether the observation is even less likely if the null hypothesis is false. Note that not having enough evidence to reject a hypothesis is not the same as having enough evidence to accept it. For more details about the differences between classical and Bayesian hypothesis testing methods, the reader is referred to Berger and Delampady [1987], Hwang et al. [1992].

1.2.3 Multi-model comparison in Bayesian setting

In some engineering applications, several representations, i.e., models, might exist with different approaches and assumptions to analyze the occurring processes. Therefore, a significant research challenge is to accurately assess competing modeling concepts and validate the corresponding computational models against an experiment or a reference solution. In the case of the multi-model comparison, the validation hypothesis is which model within the pool of available models can represent the data-generating process. The data could be observed values in the experiments or the reference data resulting

from a detailed simulation. The uncertainty concerning the choice of adequate representation of the system of interest is known as conceptual uncertainty. We may analyze this multi-model problem in a validation benchmark.

Bernardo and Smith [2009] define the following multi-model settings from a decision-theoretic perspective:

- $\mathcal{M}$-closed setting: the true data-generating model is among the finite set of candidate models,
- $\mathcal{M}$-complete: the true model exists and can be approximated but is outside the model set,
- $\mathcal{M}$-open: the true model is not within the model set and cannot even be described given the existing information.

We deal with a finite set of distinct and fully defined models in most engineering applications. These models share at least one objective, such as an SRQ in an $\mathcal{M}$-open setting, where a true data-generating model is unavailable. Thus, aiming to identify a true data-generating model is not a suitable objective. Then, dealing with conceptual uncertainty in a model ranking problem does not relate to the probability of being the true model anymore. However, the ranking may only provide the best choice in an imperfect set of models.

An objective model ranking can help us in a multi-model comparison and prevent us from an unreasonable preference for one model over its variants [Elliott and Brook, 2007]. This ranking must follow the qualitative heuristic of Occam's razor (e.g., Hutter [2007], and references therein). This principle states that a model is best among competing hypotheses if it requires the fewest assumptions; hence is the simplest model while still being consistent with the observations. This implies that, for a given amount of data, there is an optimal complexity of a model [Claeskens, 2016], which is neither too complex nor too simple. A model that follows this principle is called *parsimonious* [Angluin and Smith, 1983].

According to Schöniger [2016], model complexity could be quantified using three approaches: a) counting the number of the calibration input parameters, b) investigating the possible parameter correlations by undertaking a factor analysis, or c) accounting for data-parameter sensitivity to assess model's flexibility in prediction space instead

of in parameter space. The latter can be investigated in a Bayesian setting, where the flexibility in parameter values is considered by prior probability functions and sensitivity to data via likelihood functions. For a detailed discussion and a didactic example, the reader is referred to Schöniger et al. [2015a], Schöniger [2016].

Several multi-model frameworks are related to these model ranking methods and allow for statistical model selection and averaging (e.g., Gelman et al. [1995], Burnham and Anderson [1998]). The most prominent example might be Bayesian model selection (BMS) or averaging (BMA) [Draper, 1995, Hoeting et al., 1999, Raftery et al., 2005], in which model probabilities are used to express uncertainty between models in terms of how likely it is that a certain candidate model generated the observed data. BMS and BMA are often the first choices to deal with conceptual uncertainty over many disciplines (e.g., Trotta [2008], Faust et al. [2013], Hooten and Hobbs [2015], Schöniger [2016]). Similarly, the so-called Pseudo-BMA [Geisser and Eddy, 1979, Yao et al., 2018] is used to handle uncertainty between multiple models concerning their individual ability to predict potential future data. Likewise, model rating methods like the famous Akaike information criterion (AIC) [Akaike, 1974, 1998] serve as a basis for model selection. Other frameworks such as (Bayesian) stacking [Le and Clarke, 2017, Yao et al., 2018] allow combining model competitors in a set for predictive purposes rather than quantifying the uncertainty about one being relatively best. BMA is out of the scope of this dissertation because it aims to average or combine the predictions instead of making model comparisons solely.

In a validation benchmark problem, one also tends to identify the impact of the data size on the ranking outcome and the similarity of the models or lack thereof. The first issue answers the question: how much data are required to achieve reasonable ranking, or which level of complexity is justified given the available data? However, the similarity can be addressed by the question: how would the models be ranked given that one of them is the data-generating problem? These aspects were included in a so-called justifiability analysis introduced by Schöniger et al. [2015a] and applied to environmental applications, e.g., reactive transport models [Schäfer Rodrigues Silva et al., 2020].

1.2.4 Computational costs of Bayesian model validation

The probabilistic nature of the Bayesian techniques such as BHT, BMS requires propagating the parametric uncertainty through all competing models – i.e., a significant number of model evaluations – to reach statistical convergence. In practice, however, the computational complexity and total computational budget severely limit the number of evaluations that can be performed. This challenge renders the brute-force computation of the Bayesian metrics required for the model validation infeasible. Therefore, Bayesian estimation resorts to using cheap-to-evaluate surrogate models to approximate the actual model evaluations for computationally expensive solutions.

The primary goal here is to replace computational models with their easy-to-evaluate surrogates that replicate the behavior of the underlying physical models from a limited set of runs without sacrificing accuracy. A surrogate-assisted Bayesian analysis has been applied to many applications, including hydrology [e.g., Yoon et al., 2022, Xu et al., 2022], groundwater contamination [Jiang et al., 2021], sediment transport [e.g., Mohammadi et al., 2018, Beckers et al., 2020], processes in subsurface reservoirs [e.g., Bazargan et al., 2015, Bazargan and Christie, 2017], and subsurface flow models [e.g., Elsheikh et al., 2014].

1.3 Goals, contributions and structure

1.3.1 Goals

This dissertation aims to help modelers to perform an uncertainty-aware model validation and/or comparison. A two-stage Bayesian multi-model framework is discussed for modeling tasks where a set of models are at hand. This framework employs Bayesian hypothesis testing and Bayesian model selection and can be used to establish validation benchmarks. It is extended to a surrogate-assisted framework to make this framework applicable for computationally demanding computational models, keeping the computational costs at a reasonable level. Using surrogate representations instead of full-fidelity computational models introduces additional errors to the validation metrics. Therefore, a correction factor is introduced to compensate for the surrogate error.

I employed polynomial chaos expansions (PCE) to build a surrogate representation of the computational model. PCE is commonly used for uncertainty propagation thanks to its solid mathematical basis and ability to provide functional representations of stochastic quantities. However, the accuracy of the prediction of these surrogate models, trained with only a handful of simulations, is debatable. This argument is rooted in the fact that the surrogates do not attempt to quantify the epistemic uncertainty associated with their predictions. In this dissertation, I show how Bayesian formalism can be materialized by employing the concept of PCE to achieve better surrogates with a sparse PCE representation and account for the uncertainty in the surrogate's predictions.

I also highlight how a surrogate model using a PCE can be constructed for computationally intensive models with as few simulations as the computational budget allows. To do so, I discuss a sequential adaptive sampling strategy, in which one attempts to augment the initial design iteratively. By doing so, interesting regions in the parameter space are adequately explored. These regions are more likely to provide valuable information on the behavior of the original model responses. Using a sequential sampling strategy avoids the waste of computational resources as opposed to the so-called one-shot designs.

To my knowledge, no open-source programming package exists that deals specifically with the uncertainty-aware Bayesian multi-model comparison. There is a need for a software package that provides an automated workflow for Bayesian calibration/validation for computationally expensive models. I introduce *BayesValidRox*, an open-source, object-oriented Python package for surrogate-based sensitivity analysis, Bayesian calibration, and Bayesian comparison of computational models with a modular structure.

In the field of subsurface hydro-system modeling, many approaches could be adopted to describe the processes of interest. Comparing these model variants in light of available data can be addressed in validation benchmarks. Another goal of this dissertation is to apply the proposed uncertainty-aware validation framework to a range of topics dealing with flow and transport in porous media. However, this validation framework can be transferred to other disciplines in which models are used, such as psychology, ecology, economics, and other engineering disciplines.

1.3.2 Contributions

This dissertation aims to make the following contributions to the scientific community:

- Developing a surrogate-assisted uncertainty-aware validation framework via a Bayesian setting,
- Improving PCE prediction accuracy by exploring the Bayesian sparse learning and sequential adaptive sampling,
- Providing an open-source Python package for model calibration and validation of computational models,
- Establishing benchmarks in the field of flow and transport in porous media.

The first contribution focuses on developing a Bayesian calibration/validation framework incorporating parameter and conceptual uncertainty. Incorporating a fully Bayesian approach yields an optimal bias-variance trade-off against the experimental data and provides an integrative quantity for model validation. Additionally, to guarantee the feasibility of the Bayesian validation framework for computationally expensive models, we accelerate the computations for expensive models via model reduction techniques.

1.3.3 Structure of this dissertation

The remainder of this dissertation is structured as follows. Chapter 2 introduces components of the Bayesian validation framework. This is followed by Chapter 3, which elucidates the contribution of surrogate modeling to offset the computational cost. This chapter includes a comparison of some sparse learning strategies for aPCE and sequential sampling strategies. Chapter 4 introduces *BayesValidRox* python package and serves as a guide for the users to couple models, define input parameters' prior distributions, specify options, and post-process the results.

The remaining chapters present the framework's application to different case studies dealing with modeling flow and transport in porous media. Chapter 5 provides the application of the proposed framework to flow simulation in the fractured porous media. Chapter 6 discusses a comparison of three modeling concepts for coupling free flow and porous medium. A comparison study of the models describing the property changes

in subsurface reservoirs due to microbially-induced calcite precipitation is presented in Chapter 7. Lastly, this dissertation is concluded with a summary and outlook in Chapter 8.

2 Bayesian Validation Framework

Bayesian epistemology offers a robust framework for characterizing scientific inference since its simple concept lies in the fact that rational belief comes in degrees that can be measured in terms of probabilities. Moreover, Bayesian epistemology has resulted in the useful elucidation of notions such as confirmation. Thus, it is proven to form a viable method for data-driven validation of computer simulations and can provide a solid basis for a sound evaluation of computer simulations.

Bayesian epistemology dates back to ideas by Thomas Bayes and has been reinforced by many scientists in the twentieth century, such as Bruno de Finetti, Frank P. Ramsey, and Leonard Savage, among others. Bayesianists believe that trust comes in degrees that are measured in terms of probabilities or probability densities when dealing with characteristics with a continuous range of values. They generally tend to proceed with the following three steps. First, they formulate plausible hypotheses related to a simulation. Second, they consider rational degrees of belief in these hypotheses. Thirdly, they apply the Bayesian principles. Following these steps yields posterior probabilities that reveal the degree of trust that one should rationally invest in the hypotheses.

What are the relevant hypotheses regarding the validation of a simulation? These could be building up or losing trust in simulations. In terms of validation, the hypothesis is whether the model can satisfactorily represent the real system of interest. Moreover, for the assessment of the physical phenomena in question, several representations, i.e., models, might exist with different approaches and assumptions to analyze the occurring processes. In this case, the hypothesis is which model within the pool of available models can represent reality, i.e., observed values in the experiments. But, how can we arrive at the probabilities above? This can be achieved by updating the prior belief based on Bayesian notions. For this purpose, simulation results need to be compared with data.

In what follows, I will introduce Bayes' theorem and show how it is used for inference problems, such as calibration. Then, the validation workflow used in this work will be presented. The last section of this chapter is dedicated to the framework with a two-stage surrogate-based Bayesian multi-model comparison.

2.1 Bayesian analysis of computer simulators

Complex models are built in almost all science and engineering fields to simulate real-world systems' behavior. These models may be empirical or represent detailed scientific understandings of the real-world process. They are usually implemented in computer programs, which can run thousands of lines of code and take a fraction of a second to several days to run. I will refer to both the mathematical model and the computer program that implements it as a *simulator*. Note that the numerical implementation of the mathematical model is referred to as *computational model*.

Simulators produce output that predicts the real-world phenomenon the model estimates. However, these outputs are bound to be imperfect. Whether the simulator's output will correspond closely to the real-world quantities is uncertain. This uncertainty emerges from many sources, particularly uncertainty due to the correct value of the input parameters and uncertainty about the correctness of the model in representing the processes involved in the system under investigation. Kennedy and O'Hagan [2001] provide a more comprehensive taxonomy of the uncertainties involved in using simulators.

Many uncertainty quantification tasks in the sciences and engineering, such as calibration, require incorporating data into a model. This approach can significantly reduce the uncertainty in model predictions and is a crucial step in many applications. In the context of computational modeling, model calibration benefits from Bayes' theorem, allowing data into a model. Bayesian model calibration focuses on identifying the input parameters of a computational model to allow one to recover the observations. A typical scenario in this respect is identifying unknown properties of crucial components of a complex system based on their observed response to controlled external loads in a laboratory experiment. Through this procedure, the inferred values (and possibly the uncertainty of the estimation) can then be used to predict the response of the same

system to different external loads or even to design different systems sharing the same calibrated model component. This approach lays the foundation for the V&V under the uncertainty paradigm, commonly used in the engineering practice [Oberkampf et al., 2004, Oberkampf and Roy, 2010, Hu and Orient, 2016]. The purpose of this section is to introduce Bayes' theorem and show how it can be used for statistical modeling. This also establishes a rigorous framework for developing uncertainty-aware validation benchmarks. Moreover, this section reviews different approaches to the Bayesian inverse problems, i.e., how the possible values of parameters can be inferred.

2.1.1 Bayes' theorem

The Bayesian approach to validation extensively exploits Bayes' theorem. This theorem is a combination of traditional probabilities and statistics. Let us assume A and B are two events. The conditional probability of event A given that we know that the event B has occurred, $P(A \mid B)$, can be cast as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \tag{2.1}$$

where $P(B)$ denotes the probability of event B, which poses a positive value, and $P(A \cap B)$ signifies the probability that both A and B occurred. Given that $P(A \cap B)$ is equal to $P(B \cap A)$, Equation (2.1) can be recast as follows:

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A). \tag{2.2}$$

Bayes' theorem can simply be obtained by dividing Equation (2.2) by $P(B)$, which gives:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2.3}$$

Bayes' theorem, as expressed in Equation (2.3), connects the conditional probability $P(B \mid A)$ to the other conditional probability $P(A \mid B)$. $P(A)$ is the prior probability and $P(A \mid B)$ stands for posterior probability, which are two basic concepts specific to Bayesian methods. While the posterior is the conditional probability based on event B, the prior is the unconditioned probability, which is used to integrate expert knowledge or previous experience into analyses.

2.1.2 Statistical modeling via Bayes' theorem

Bayesian statistics uses Bayes' theorem, described earlier, to fit a statistical model to the problem at hand. This is achieved by updating the prior knowledge on hyper-parameters $\boldsymbol{\theta}$, defined by the conditional probability $\boldsymbol{\theta} \sim P(\boldsymbol{\theta})$, with possibly few observation data points. These hyper-parameters are treated as random variables, whose subjective definition must disclose the available information *before* any measurement of the quantity of interest, $\mathcal{Y}$, is performed.

Bayes' theorem in the context of statistical inference can be recast as the following:

$$P(\boldsymbol{\theta} \mid \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{Y})}, \tag{2.4}$$

where $P(\boldsymbol{\theta} \mid \mathcal{Y})$ denotes the posterior distribution of the hyper-parameters, $p(\mathcal{Y} \mid \boldsymbol{\theta})$ the likelihood function and $P(\mathcal{Y})$ is the probability of the data or marginal likelihood.

Likelihood function Inverse Bayesian problems require three ingredients: a computational forward model $\mathcal{M}$, a set of input parameters $\boldsymbol{\theta} \in \Theta$ and a set of observed measurement data $\mathcal{Y}$. The forward model $\mathcal{M}$ is a computational model that receives a set of parameters $\boldsymbol{\theta}$ and generates outputs. This computational model uses computer programs to simulate complex systems using an algorithmic or mechanistic approach. Still, they involve simplifications of the real system and are often in error due to the lack of proper understanding of governing phenomena. Therefore, no models can precisely simulate reality; thus, a discrepancy always exists. As a result, the observed measurements $\mathcal{Y}$ can be related to the model by introducing a discrepancy term $\epsilon \in \mathbb{R}^{N_{out}}$, which represents the effects of measurement error and model inaccuracy. This relation can be cast as follows:

$$\mathcal{Y} = \mathcal{M}(\boldsymbol{\theta}) + \epsilon. \tag{2.5}$$

A common choice for the additive discrepancy term in engineering disciplines is a simple Gaussian distribution. Other distribution types are also used in a more general setting [Schoups and Vrugt, 2010]. This choice of ϵ indicates that the discrepancy is a random value from a normal distribution $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, which is centered at zero and has a covariance matrix of $\boldsymbol{\Sigma}$. For the sake of simplicity, the distribution is chosen to be centered at zero. However, one can include a model discrepancy function as a bias

term to account for systematic model error. Ling et al. [2014] provides a good guideline for selecting an appropriate prior form for the model discrepancy. This approach was used in many engineering applications, such as Bayesian calibration of urban drainage model Nagel et al. [2020] and heat transfer models for fire insulation panels [Wagner et al., 2020].

According to the discussion above, a measurement point $y_i \in \mathcal{Y}$ is a realization of a multivariate Gaussian (normal) distribution $\mathcal{N}(y_i \mid \mathcal{M}(\boldsymbol{\theta}), \boldsymbol{\Sigma})$, which is centered at $\mathcal{M}(\boldsymbol{\theta})$ and have a covariance matrix $\boldsymbol{\Sigma}$ [Prince, 2012]. Assuming that a measured data set of $\mathcal{Y} = (y_1, \cdots, y_N)^T$ with the independent realization is available, the probability of observing the data can be defined by the likelihood function as the following:

$$\begin{aligned} p(\mathcal{Y} \mid \boldsymbol{\theta}) &:= \prod_{i=1}^{N} \mathcal{N}(y_i \mid \mathcal{M}(\boldsymbol{\theta}), \boldsymbol{\Sigma}) \\ &= \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathcal{M}(\boldsymbol{\theta}) - \mathcal{Y})^T \boldsymbol{\Sigma}^{-1} (\mathcal{M}(\boldsymbol{\theta}) - \mathcal{Y})\right). \end{aligned} \tag{2.6}$$

Posterior distribution By combining the prior and the likelihood in Equation (2.6), the posterior distribution in Equation (2.4) can be computed by

$$P(\boldsymbol{\theta} \mid \mathcal{Y}) = \frac{1}{P(\mathcal{Y})} P(\boldsymbol{\theta}) \cdot \prod_{i=1}^{N} \mathcal{N}(y_i \mid \mathcal{M}(\boldsymbol{\theta}), \boldsymbol{\Sigma}). \tag{2.7}$$

After conditioning on the observed data, the posterior distribution summarizes the inferred information regarding the hyper-parameters $\boldsymbol{\theta}$. However, the practical computation of the posterior distribution $P(\boldsymbol{\theta} \mid \mathcal{Y})$ is not trivial. Analytical expressions exist only for particular choices of the prior distribution $P(\boldsymbol{\theta})$. One approach for more involved cases is to use sampling methods, such as Monte Carlo or Markov Chain Monte Carlo, to approximate the posterior distribution. The posterior distribution approximation is a task of inference and will be detailed in Section 2.1.3.

Marginal likelihood The denominator in Equation (2.4), $P(\mathcal{Y})$, is a normalization factor, which ensures that the posterior probability sums up to one. It is also known as Bayesian model evidence (BME), or prior predictive since it quantifies the likelihood

of the observed data based on the prior distribution of the parameters. It can take the following form:

$$P(\mathcal{Y}) = \int_{\Theta} p(\mathcal{Y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{2.8}$$

The marginal likelihood is not analytically tractable and must be approximated via numerical methods. Numerous methods have been proposed in the literature, such as the Laplace approximation, harmonic mean estimator, annealed importance sampling, nested sampling, path sampling, bridge sampling, parallel tempering with thermodynamic integration, and reversible jump MCMC. For comprehensive studies about different methods for the marginal likelihood approximation, see Han and Carlin [2000], Chen et al. [2012], Robert and Wraith [2009], Marin and Robert [2009], Friel and Wyse [2012]. This study employs the most straightforward method for estimating marginal likelihood, which is the naive Monte Carlo method [Hammersley, 2013, Raftery and Banfield, 1991]. This method relies on the idea that the marginal likelihood can be written as an expected value with respect to the prior distribution, i.e., $P(\mathcal{Y}) = \mathbb{E}_{P(\boldsymbol{\theta})}\left[p(\mathcal{Y} \mid \boldsymbol{\theta})\right]$. This expected value can be approximated by evaluating the likelihood in (2.6) for N_{MC} samples, drawn from the prior, $P(\boldsymbol{\theta})$ and averaging the resulting likelihoods. This procedure yields a naive Monte Carlo estimator $\hat{P}(\mathcal{Y})$, which reads

$$\hat{P}(\mathcal{Y}) = \frac{1}{N_{\mathrm{MC}}} \sum_{i=1}^{N} p\left(\mathcal{Y} \mid \tilde{\boldsymbol{\theta}}_i\right), \quad \tilde{\boldsymbol{\theta}}_i \sim p(\boldsymbol{\theta}). \tag{2.9}$$

Predictive distributions One advantage of using the Bayesian inference in the task of model validation is that the uncertainty on hyper-paramters $\boldsymbol{\theta}$ can be incorporated into prior and posterior evaluations of quantities of interest y by so-called *predictive distributions*. The following expression gives the prior predictive distribution:

$$P'_{pred}(y) := \int_{\Theta} p(y \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.10}$$

which is the average of the conditional distribution of the prior distribution over the prior distribution $P(\boldsymbol{\theta})$. Similarly, the posterior predictive distribution $P''_{pred}(y \mid \mathcal{Y})$ can be computed by averaging the conditional distribution $P(y \mid \boldsymbol{\theta})$ over the posterior

parameter distribution $P(\theta \mid \mathcal{Y})$, defined in Equation (2.4). It reads

$$P''_{pred}(y) := \int_{\Theta} p(y \mid \boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \mathcal{Y}) d\boldsymbol{\theta} = \frac{1}{P(\mathcal{Y})} \int_{\Theta} p(y \mid \boldsymbol{\theta}) p(\mathcal{Y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{2.11}$$

2.1.3 Bayesian inference

The posterior distribution $P(\boldsymbol{\theta} \mid \mathcal{Y})$ in Equation (2.4) provides a summary of inferred information concerning the uncertain parameters after updating the prior knowledge in light of the observed data. Its practical computation, however, is a non-trivial task. There exist only ad-hoc solutions for particular choices of prior and likelihood distributions. However, for more involved cases, the posterior distribution can only be numerically approximated using Monte Carlo sampling methods. In what follows, two methods for approximating the posterior distribution $P(\boldsymbol{\theta} \mid \mathcal{Y})$ are introduced: rejection sampling and Markov Chain Monte Carlo. Both methods relate to the general field of Monte Carlo techniques and use a proxy distribution to approximate the target posterior distribution.

Rejection sampling

The rejection sampling algorithm draws samples from difficult-to-approximate distributions by using a proxy distribution. This method is also known as the acceptance-rejection or accept-reject method. Using this method, sampling values $\boldsymbol{\theta}$ from a target distribution with an arbitrary probability density function $f(\boldsymbol{\theta})$ are generated via a proposal distribution $\boldsymbol{\theta}^{(*)}$ with a probability density $g(\boldsymbol{\theta}^{(*)})$. The idea behind rejection sampling is that one can generate a sample value $\boldsymbol{\theta}^{(*)}$ from g instead of sampling from f and accept the sample with probability $f(\boldsymbol{\theta})/(Mg(\boldsymbol{\theta}))$, repeating the draws from $g(\boldsymbol{\theta}^{(*)})$ until a value is accepted. M is a constant, finite bound on the likelihood ratio $f(\boldsymbol{\theta})/g(\boldsymbol{\theta})$, satisfying $1 < M < \infty$ over the support of $\boldsymbol{\theta}$; in other words, M must meet $f(\boldsymbol{\theta}) \leq Mg(\boldsymbol{\theta})$ for all values of θ. This can be the maximum of $p(\boldsymbol{\theta} \mid \mathcal{Y})$. It is worth mentioning that this requires that the support of $\boldsymbol{\theta}^{(*)}$ must include the support of θ—in other words, $g(\boldsymbol{\theta}) > 0$ whenever $f(\boldsymbol{\theta}) > 0$.

Suppose that $g(\boldsymbol{\theta})$ is a positive function for all $\boldsymbol{\theta}$ for which $p(\boldsymbol{\theta} \mid \mathcal{Y}) > 0$. The following steps should be taken:

1. Draw sample $\boldsymbol{\theta}^{(*)}$ randomly from the probability proportional to $g(\boldsymbol{\theta})$.
2. Draw sample u from $\mathcal{U}(0, 1)$, and
3. Check whether $p(\boldsymbol{\theta}^{(*)} \mid \mathcal{Y})/(M \cdot g(\boldsymbol{\theta}^{(*)})) \leq u$.
    - If holds, accept $\boldsymbol{\theta}^{(*)}$ as a sample drawn from $p(\boldsymbol{\theta} \mid \mathcal{Y})$,
    - if not, reject the value of $\boldsymbol{\theta}^{(*)}$ and return to step 1.

The accepted samples $\boldsymbol{\theta}$ correspond to drawn samples from $p(\boldsymbol{\theta} \mid \mathcal{Y})$.

Despite a simple algorithm, the rejection sampling method has some challenges and drawbacks. As per of challenges, the selection of the appropriate proposal function $g(\boldsymbol{\theta})$ and finding its scaling constant M are not trivial. Rejection sampling can lead to many unwanted samples being taken if the function being sampled is highly concentrated in a particular region, for example, a function with a spike at some location [Gilks and Wild, 1992]. Generally, this method is inefficient especially in higher dimensions, because most samples are rejected for high-dimensional complex distribution [Law, 2019].

MCMC: Markov Chain Monte Carlo sampling

As indicated earlier, there is no closed-form solution for the posterior distribution for any prior and likelihood distribution choices. Therefore, Monte Carlo methods are commonly used to approximate the posterior distribution $P(\boldsymbol{\theta} \mid \mathcal{Y})$ in Equation (2.4). One common technique for performing such inference tasks is to employ the MCMC methods [Robert et al., 1999, Liu and Liu, 2001]. These methods were first proposed by Metropolis et al. [1953] and extended by Hastings [1970], but their impact in the field of Statistics was not felt until much later. Gelfand and Smith [1990], however, underscored that it was possible to produce a Markov chain with a specified stationary distribution f.

The main goal of MCMC methods is to draw samples $\boldsymbol{\theta}$ from the posterior probability density $P(\boldsymbol{\theta} \mid \mathcal{Y})$ by generating a random walk in the parameter space. Note that MCMC, similar to rejection sampling, uses a proxy distribution to approximate the target posterior distribution. Moreover, there is no need to compute the normalization

factor $P(\mathcal{Y})$ once the goal is to perform a parameter inference. Therefore, the non-normalized posterior probabilities can be computed by

$$P(\boldsymbol{\theta} \mid \mathcal{Y}) \propto p(\mathcal{Y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}). \tag{2.12}$$

MCMC constructs a Markov chain $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots\}$ in the parameter space Θ with an invariant distribution that is equivalent to the target posterior distribution. Markov chains can be defined by their transition probability $\kappa(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^{(k)})$ from the step $\boldsymbol{\theta}^{(k)}$ of the chain at iteration k to the step $\boldsymbol{\theta}^{(k+1)}$ at the subsequent iteration $k+1$. Moreover, the so-called detailed balance condition for a Markov chain reads

$$p(\boldsymbol{\theta}^{(k)} \mid \mathcal{Y}) \kappa(\boldsymbol{\theta}^{(k+1)} \mid \theta^{(k)}) = p(\boldsymbol{\theta}^{(k+1)} \mid \mathcal{Y}) \kappa(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k+1)}), \tag{2.13}$$

which ensures the reversibility of the Markov chains and states that the probability of being at $\boldsymbol{\theta}^{(k)}$ and moving to $\boldsymbol{\theta}^{(k+1)}$ is equal to the probability of being at $\boldsymbol{\theta}^{(k+1)}$ and moving to $\boldsymbol{\theta}^{(k)}$. If the specified transition probability meets the detailed balance condition, the posterior is believed to be the invariant distribution of the Markov chains.

By integrating the condition in (2.13) over $d\boldsymbol{\theta}^{(k)}$ which reads

$$p(\boldsymbol{\theta}^{(k+1)} \mid \mathcal{Y}) = \int_{\Theta} p(\boldsymbol{\theta}^{(k)} \mid \mathcal{Y}) \kappa(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^{(k)}) d\boldsymbol{\theta}^{(k)}, \tag{2.14}$$

one can show that the transition distribution of the Markov chain at its equilibrium state is equal to the posterior distribution. One needs to record states from the chains to obtain samples of the desired posterior distribution.

A Metropolis-Hasting algorithm (MH) [Metropolis et al., 1953, Hastings, 1970] is one method of the Monte Carlo class that ensures that the state of the detailed balance condition in (2.13) is met. By generating a sequence of sample values, the algorithm will create a distribution of values that approaches the desired distribution as more and more samples are produced. The algorithm selects a candidate for the next sample value at each iteration based on the current sample value, forming a Markov chain. The candidate is then either accepted with some probability or rejected. In the former case, the candidate value is used in the next iteration, and in the latter, the candidate sample is excluded, and the current value is reused in the next iteration. To calculate

the probability of acceptance, the current and candidate sample values of the proposal function are compared with respect to the desired posterior distribution.

Let us assume that a chain is initialized at a particular seed point $\boldsymbol{\theta}^{(0)}$, which belongs to the parameter space Θ. Then, we can draw a candidate point $\boldsymbol{\theta}^{(\star)}$ from a proposal distribution $p(\boldsymbol{\theta}^{(\star)} \mid \boldsymbol{\theta}^{(k)})$ at the iteration k from the current point $\boldsymbol{\theta}^{(k)}$. In practice, to accept or reject the candidate $\boldsymbol{\theta}^{(\star)}$, we first select a random value u from a uniform distribution $u \sim \mathcal{U}(0, 1)$. Then, we compare u with the ratio α, which reads

$$\alpha\left(\boldsymbol{\theta}^{(\star)} \mid \boldsymbol{\theta}^{(k)}\right) = \min\left\{1, \frac{p(\boldsymbol{\theta}^{(\star)} \mid \mathcal{Y})\, p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(\star)})}{p(\boldsymbol{\theta}^{(k)} \mid \mathcal{Y})\, p(\boldsymbol{\theta}^{(\star)} \mid \boldsymbol{\theta}^{(k)})}\right\}. \tag{2.15}$$

The candidate $\boldsymbol{\theta}^{(\star)}$ is accepted with the probability α, if $\alpha \geq u$. Otherwise, $\boldsymbol{\theta}^{(\star)}$ is rejected and the proposal at the iteration $k+1$ is set to the current point ($\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)}$). In Equation (2.15), the proposal distribution is denoted by $p(\boldsymbol{\theta}^{(\star)} \mid \boldsymbol{\theta}^{(k)})$. A commonly used proposal distribution is the Gaussian distribution $p(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\Sigma}_p)$. This proposal is symmetric and centered at the value of the current iteration, $\boldsymbol{\theta}^{(k)}$ with a covariance matrix $\boldsymbol{\Sigma}_p$. Using a symmetric proposal distribution, a candidate is accepted if $p(\boldsymbol{\theta}^{(\star)} \mid \mathcal{Y}) \geq p(\boldsymbol{\theta}^{(k)} \mid \mathcal{Y})$, that means that the candidate more likely belongs to the posterior distribution than the current candidate $\boldsymbol{\theta}^{(k)}$. However, if $p(\boldsymbol{\theta}^{(\star)} \mid \mathcal{Y}) < p(\boldsymbol{\theta}^{(k)} \mid \mathcal{Y})$, the proposed candidate is accepted only with the probability $\alpha = p(\boldsymbol{\theta}^{(\star)} \mid \mathcal{Y})/p(\boldsymbol{\theta}^{(\star)} \mid \boldsymbol{\theta}^{(k)})$.

Wagner et al. [2021] argue that Metropolis-Hastings algorithms have a practical weakness in that they require a proposal distribution $p(\boldsymbol{\theta}^{(\star)} \mid \boldsymbol{\theta}^{(k)})$. In principle, this distribution has to resemble that of the posterior distribution. However, in most applications, the posterior shape is seldom known *a priori*. Additionally, a poorly chosen proposal distribution significantly deteriorates the MCMC algorithm's performance, resulting in the algorithm failing since no new candidates are accepted. This phenomenon is typically more prominent in high dimensions with strongly correlated posterior distributions. Haario et al. [2001] propose an extension of MH, known as *adaptive Metropolis algorithm* (AM), in which the Gaussian proposal distribution of MH is adapted during sampling based on the previously generated samples. The MH and AM follow a random walk concept in finding proposals. Duane et al. [1987] introduce *Hamiltonian Monte Carlo algorithms* (HMC) that exploits the gradient of the posterior distribution using Hamiltonian dynamics to construct a Markov chain. For more details, see Neal et al.

[2011], Nagel and Sudret [2016].

MCMC algorithms often do not perform well when the target distribution (i.e., posterior distribution) exhibits a strong correlation between the parameters. A considerable amount of tuning is typically needed to improve the performance of these algorithms. To alleviate this problem, Goodman and Weare [2010] present the affine invariant ensemble sampler (AIES). As opposed to the classical MCMC algorithms, AIES can draw samples from both types of distributions with or without correlation without explicitly requiring the affine transformation of the target distribution [Wagner et al., 2021].

The AIES algorithm runs an ensemble of L Markov chains, also known as *walkers* $\left\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_L\right\}$. The current location of the points $\boldsymbol{\theta}_i^{(k)}$ is updated by random selection of a conjugate walker $\boldsymbol{\theta}_j^{(k)}$ from the walker sets, excluding the current i-th walker ($i \neq j$). A so-called *stretch move* can be used to generate proposed samples that ensure the affine invariant property of the algorithm. The candidate is proposed by

$$\boldsymbol{\theta}_i^{(\star)} = \boldsymbol{\theta}_i^{(k)} + Z\left(\boldsymbol{\theta}_j^{(\tilde{k})} - \boldsymbol{\theta}_i^{(k)}\right), \tag{2.16}$$

where $\tilde{k} = k+1$ if $j < i$ and $\tilde{k} = k$, otherwise. Z is randomly drawn from the probability distribution function $p(z \mid a)$ which reads

$$p(z \mid a) = \begin{cases} \frac{1}{\sqrt{2}\left(2\sqrt{a} - \frac{2}{\sqrt{a}}\right)} & \text{if } z \in [1/a, a] \\ 0 & \text{otherwise,} \end{cases} \tag{2.17}$$

where a is the tuning parameter, which is greater than one.

For the i-th walker in the ensemble, the candidate point $\boldsymbol{\theta}_i^{(\star)}$ is accepted with the following probability as the new location

$$\alpha\left(\boldsymbol{\theta}_i^{(\star)}, \boldsymbol{\theta}_i^{(k)}, z\right) = \min\left\{1, z^{M-1}\frac{p\left(\boldsymbol{\theta}_i^{(\star)} \mid \mathcal{Y}\right)}{p\left(\boldsymbol{\theta}_i^{(k)} \mid \mathcal{Y}\right)}\right\}. \tag{2.18}$$

Similar to other MCMC algorithms, α is then compared with a randomly drawn value $u \sim \mathcal{U}(0,1)$. This procedure will be repeated for all L walkers in the ensemble. As opposed to MCMC algorithms, AIES has only one scalar tuning parameter a, whose value can be set to 2, suggested by Goodman and Weare [2010], Allison and Dunkley

[2013] and Wicaksono [2018]. In this dissertation, we use the Python implementation of AIES, *emcee* [Foreman-Mackey et al., 2019].

One of the significant challenges for MCMC users is determining when the convergence is reached, i.e., when it is safe to terminate sampling and use the samples to estimate the characteristics of the poster distribution. Most MCMC users address this problem by performing convergence diagnostics. Cowles and Carlin [1996] provide an exhaustive review of thirteen convergence monitoring strategies for convergence diagnostics. They recommend that a combination of tools can be beneficial for evaluating the performance and convergence of the MCMC sampler. The proposed framework benefits from two convergence methods: Robin & Gelman method and integrated autocorrelation time, whose definitions can be found in Appendix A.

2.2 Validation workflow

This section introduces the workflow of the Bayesian validation framework. This workflow aims to simplify the framework implementation for an easy-to-follow validation task. In the very first step, the goal of validation needs to be clearly defined by determining the essential physical phenomena to be investigated. Then, models are designed to describe these physical phenomena and expose the parameters in these models for the task of uncertainty quantification. In addition to the parameters, the SRQ must be selected for the study. These are the quantities based on which the model simulations and the observed reality in the form of a controlled experiment are compared to.

Once all parameters and SRQs are available, the next step involves constructing the parameter probability density functions (PDFs). The parameter PDFs are usually constructed based on expert opinion. This opinion is based on the experience of someone knowledgeable about the field. Most often, they provide a uniform parameter distribution with estimates of the lower and the upper bounds. Since running simulations using these parameters influence curve fits to experimental data, this approach is highly error-prone unless the expert has detailed knowledge of the experiment. To avoid this subjective judgment in the validation of models, Jiang et al. [2019] argue that Bayesian validation methods must be fully integrated with the model parameter calibration to yield more accurate prediction results in light of both data and modeling uncertainties.

For the task of parameter calibration in light of uncertainties, Bayesian calibration can be employed to complement the very subjective approach, such as expert opinion, with a precise and mathematically sound approach like Bayesian calibration, outlined in Section 2.1.3. It is thus easier to defend the parameter distribution since it and all implied uncertainties are precisely defined, and this is a process that a quantification analyst can repeat. This step yields a so-called *posterior* parameter distribution, which is an updated belief on the parameter distribution after comparing with data.

The posterior parameter distribution is then used as the parameter distribution for the validation phase. In validation, the statistical hypothesis testing method can be used to investigate the validity of a computational model in light of data obtained from an experiment or a detailed reference model. This data must be different from what has been used already in the Bayesian inference step. In a multi-model setting, where more than one model is available to simulate SRQs, one can additionally analyze the models via a Bayesian multi-model comparison. The benefit of this comparison is twofold. The models are compared to each other based on their predictive abilities in reproducing the reference. Moreover, the models can also be compared to each other, revealing their similarities.

The rest of this section is structured as follows: first, a description of Bayesian hypothesis testing is presented. This section will be followed by a description of the Bayesian multi-model comparison.

2.2.1 Bayesian hypothesis testing

As mentioned in Section 1.2.2, statistical hypothesis testing is commonly used to perform uncertainty-aware validation of computational models. Two types of hypothesis testing methods, namely classical and Bayesian methods, can be employed to develop model validation metrics. Berger and Sellke [1987], Edwards et al. [1963] and Leamer [1978] argue that the results of the classical hypothesis testing are hard to interpret and often misleading. Alternatively, the Bayesian approach to hypothesis testing incorporates assumptions on the prior distribution under the alternative hypothesis [Jeffreys, 1961]. Therefore, we will use BHT in the proposed validation framework.

The key component in BHT is the Bayes factor. It can be used to compare the performance of two candidate models or to compare many models in pairs based on their ability to reproduce the observed data. The review paper of Kass and Raftery [1995] provides an extensive description in the context of practical applications. The authors define the Bayes factor as the ratio of BME for two alternative models.

The posterior probability $P(\mathcal{M} \mid \mathcal{Y})$ of model $\mathcal{M}$ given data $\mathcal{Y}$, is given by Bayes' theorem as

$$P(\mathcal{M} \mid \mathcal{Y}) = \frac{P(\mathcal{Y} \mid \mathcal{M})P(\mathcal{M})}{P(\mathcal{Y})}, \tag{2.19}$$

where $P(\mathcal{M})$ is the prior model weight and $P(\mathcal{Y} \mid \mathcal{M})$ denotes the probability that some data is generated by the model $\mathcal{M}$ and can be computed by BME.

Let $\mathcal{M}_k$ and $\mathcal{M}_l$ be two models under investigation, parameterized by model parameter vectors $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_l$. Also, we denote their prior probabilities to be accepted with $P(\mathcal{M}_k)$ and $P(\mathcal{M}_l)$, respectively. In a model comparison problem, we aim to compare two models based on their ability to reproduce the observed data $\mathcal{Y}$. The relative posterior probabilities of two hypotheses when an observation is made can be cast as:

$$\frac{P(\mathcal{M}_k \mid \mathcal{Y})}{P(\mathcal{M}_l \mid \mathcal{Y})} = \left[\frac{P(\mathcal{Y} \mid \mathcal{M}_k)}{P(\mathcal{Y} \mid \mathcal{M}_l)}\right] \left[\frac{P(\mathcal{M}_k)}{P(\mathcal{M}_l)}\right]. \tag{2.20}$$

The first set of square brackets on the right-hand side is data dependent and is called the Bayes factor (BF), also known as (weighted) likelihood ratio of $\mathcal{M}_k$ to $\mathcal{M}_l$. By rearranging Equation (2.20) and the Bayes factor reads:

$$\mathrm{BF}(\mathcal{M}_k, \mathcal{M}_l) = \left[\frac{P(\mathcal{M}_k \mid \mathcal{Y})}{P(\mathcal{M}_l \mid \mathcal{Y})}\right] \left[\frac{P(\mathcal{M}_l)}{P(\mathcal{M}_k)}\right] = \frac{P(\mathcal{Y} \mid M_k)}{P(\mathcal{Y} \mid M_l)}. \tag{2.21}$$

The Bayes factor, $\mathrm{BF}(\mathcal{M}_k, \mathcal{M}_l)$, can be regarded as a measure of significance in BHT and can be interpreted as the ratio between the posterior and prior odds of model M_k being the more plausible one compared to the alternative model $\mathcal{M}_l$. In other words, one can say data is to favor $\mathcal{M}_k$ over $\mathcal{M}_l$. It quantifies the evidence (literally, as in Bayesian model evidence) of the hypothesis $\mathcal{M}_k$ against the null hypothesis $\mathcal{M}_l$.

Jeffreys provided a rule of thumb in his book, Theory of probability [Jeffreys, 1961] for the interpretation of the Bayes factor for model acceptance on $\log_{10}$-scale in units of $1/2$. According to him, $\log_{10}\mathrm{BF} > 1/2$ expresses "substantial" evidence in favor of $\mathcal{M}_k$

over $\mathcal{M}_l$. A "strong" evidence can be represented by $\log_{10}\text{BF} > 1$ and $\log_{10}\text{BF} > 2$ shows a "decisive" evidence. Note that one has increasing confidence in the hypothesis $\mathcal{M}_k$ over its counterpart $\mathcal{M}_l$, as the Bayes factor increases. Kass and Raftery [1995] proposed widely used alternative grades of evidence that are summarized in Table 2.1.

Table 2.1: Interpretation of Bayes Factor according to Jeffreys [1961]

BF	Interpretation
1 - 3	anecdotal evidence
3 - 10	substantial evidence
10 - 100	strong evidence
> 100	decisive evidence

Following this suggestion, a Bayes factor which lies between 1 and 3 indicates "anecdotal" evidence in favor of $\mathcal{M}_k$ that is "not worth more than a bare mention", a factor of up to 10 represents "substantial" evidence, and a factor between 10 and 100 can be regarded "strong" evidence. Finally, a Bayes factor greater than 100 admits "decisive" evidence, i.e., it can be used as a threshold to reject models based on poor performance in comparison to the best performing model in the set.

2.2.2 Bayesian multi-Model comparison

In a multi-model setting, the individual models are compared against each other. The benefits of this comparison are twofold. First, this evaluates their strengths and weaknesses. Second, their predictive ability is assessed. The model comparison can be performed with the help of model weights. These weights are determined according to the predictive capability of the models and often a penalty for complexity, thus favoring models with greater robustness [Schöniger, 2016].

Various strategies have been suggested in the literature to compare alternative computational models with possibly distinct conceptual models. The most prominent example is BMS or BMA proposed by Draper [1995] and Hoeting et al. [1999]. BMS is a formal statistical approach that allows for comparing alternative conceptual models, testing their adequacy, combining their predictions into a more robust output estimate, and quantifying the contribution of conceptual uncertainty to the overall prediction uncertainty. The BMS method is grounded on Bayes' theorem, which, as mentioned

earlier, combines a prior belief about the efficacy of each model with its performance in replicating a common measurement data set. It returns model weights representing posterior probabilities for each model to be the most appropriate from the set of proposed competing models. Additionally, the computed weights can provide a ranking and a quantitative comparison of the competing models.

BMS closely follows the principle of parsimony or Occam's razor [Angluin and Smith, 1983], in that, the posterior model weights offer a compromise between model complexity and goodness of fit, also known as the bias-variance trade-off [Geman et al., 1992]. That means, if competing models make equally likely predictions, a less complex model is given more weight.

Bayesian Model Selection

Let us consider that N_m plausible, competing models $\mathcal{M}_k$ are available. The posterior predictive distribution of a quantity of interest $\boldsymbol{\theta}$ in Equation (2.11) given the vector of observed data $\mathcal{Y}$ can be expressed as:

$$P''_{\text{pred}}(\boldsymbol{\theta} \mid \mathcal{Y}) := \sum_{k=1}^{N_m} p(\boldsymbol{\theta} \mid \mathcal{Y}, \mathcal{M}_k) P(\mathcal{M}_k \mid \mathcal{Y}), \tag{2.22}$$

with $P(\mathcal{M}_k \mid \mathcal{Y})$ being discrete posterior model probabilities or weights, which describe the plausibility of the model $\mathcal{M}_k$ to be the data-generating model after observing the data. The weights can be interpreted as the Bayesian probability of the individual models to be the best representation of the system from the pool of competing models. The model weights (posterior probabilities of models) are given by Bayes' theorem, which can be recast for a set of N_m competing models $\mathcal{M}_k$ as:

$$P(\mathcal{M}_k \mid \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \mathcal{M}_k) P(\mathcal{M}_k)}{\sum_{i=1}^{N_m} p(\mathcal{Y} \mid \mathcal{M}_i) P(\mathcal{M}_i)}, \tag{2.23}$$

where $P(\mathcal{M}_k)$ denotes the prior probability, also known as the subjective model credibility that model $\mathcal{M}_k$ could be the most plausible model in the set of models *before* any comparison with observed data have been made. Hoeting et al. [1999] proposed that a "reasonable, neutral choice" could be equally likely priors, i.e., $P(\mathcal{M}_k) = 1/N_m$, in case of paucity of prior knowledge regarding the merit of the different models under

study. However, one may decide to assign lower prior probabilities to models with many parameters than to models with few parameters [Wilson et al., 2010, Consonni et al., 2018].

The denominator in Equation (2.23) is the normalizing constant of the posterior distribution of the models and can simply be obtained by determination of the individual weights. It ensures that the posterior model weights sum up to one. In Equation (2.23), the term $p(\mathcal{Y} \mid \mathcal{M}_k)$ represents BME, as introduced in Section 2.1.2. The BME of model $\mathcal{M}_k$ can be estimated by integration over the full parameter space Θ_k, which is known as Bayesian integral by Kass and Raftery [1995], and takes the following form:

$$p(\mathcal{Y} \mid \mathcal{M}_k) = \int_{\Theta_k} p(\mathcal{Y} \mid \mathcal{M}_k, \boldsymbol{\theta}_k) P(\boldsymbol{\theta}_k \mid \mathcal{M}_k) d\boldsymbol{\theta}_k, \tag{2.24}$$

with $\boldsymbol{\theta}_k$ being the parameter vector of model $\mathcal{M}_k$ with the dimension of $N_{p,k}$. Θ_k denotes the parameter space of model $\mathcal{M}_k$, and $P(\boldsymbol{\theta}_k \mid \mathcal{M}_k)$ is the corresponding prior distribution. In Equation (2.24), the likelihood or probability of the parameter set θ_k of model $\mathcal{M}_k$ to have generated the measurement data set is represented by $p(\mathcal{Y} \mid \mathcal{M}_k, \theta_k)$. For more information about the computation of this term, see Section 2.1.2. For more details on the properties of BME and a comparison of available techniques to evaluate this term, the reader is referred to Schöniger et al. [2014].

BMS does not necessarily involve identifying the *true* model, but rather identifying the model that one should have the strongest belief in, given the assumptions made [Hinne et al., 2020]. This belief is conditioned on the data as well as the collection of models considered. For a more in-depth discussion of this issue, see Gronau and Wagenmakers [2019a,b]. A practical limitation of BMS is that the approach requires the specification of prior distributions on the parameters of each model and the models' distribution. Identifying appropriate prior distributions is not always straightforward, and even uniform or vague priors may substantially influence the outcome. See Wagenmakers et al. [2018] for more discussion on this topic. As for models' prior parameters, we seek to use the posterior of the Bayesian calibration as the prior for BMS in the validation.

Model Justifiability Analysis

Compared to other multi-model approaches, BMS follows the principle of parsimony (Occam's razor). This leads to an optimal trade-off between goodness-of-fit in light of limited data during calibration, and it helps identify the robustness of future predictions. In the unlikely event that the "true" model is in the chosen set of models (which is never the case), then BMS will tend to prefer simpler models over the true model up to a certain point where the available data cannot justify the true complexity. In contrast, if there are infinite data points, BMS will detect the true model with absolute certainty, no matter how simple or complex it may be. This behavior is an essential characteristic of model selection frameworks, but many other model ranking techniques, especially those with the information criteria, such as AIC or BIC, cannot guarantee it [Burnham and Anderson, 2004, Schöniger, 2016].

To investigate this trade-off, Schöniger et al. [2015a] introduced a so-called *justifiability analysis*. This analysis aims at addressing the question how much data are needed to reasonably calibrate a highly complex model, or which level of complexity is justified given the available data? In this framework, one tests the modeling alternatives against each other, and tries to address the following question: How would the models be ranked given that one of them is actually the data generating model? In BMA terms, this is referred to as a justifiability analysis because it evaluates whether the level of complexity can be justified by the amount and type of data available. To identify the most efficient measurement design, one can run the proposed analysis before the actual data collection. Schöniger et al. [2015a] demonstrate this analysis in a case of model selection between groundwater models with different complexity. Schäfer Rodrigues Silva et al. [2020] used this method in determining how similar the simplified reactive transport models are to a reference model.

The Concept of model justifiability Assume that a set of models exists with different levels of complexity. Each of the models is used to generate a number of synthetic data sets, i.e., samples from the predictive distribution of a model are now used as data sets. Then, one needs to perform a standard BMS analysis to obtain the model weights. These weights are averaged over all generated synthetic data set that were generated from a specific model. As a result, this procedure yields as many sets of model weights

as the number of compared models. These averaged weights can then be reported in a form of a so-called *confusion matrix*.

Using the model confusion matrix, we can determine the maximum complexity that might be justified based on the experimental setup. If a given model receives the highest model weight in the set when it generates data, the complexity of the model is considered to be justified. Having a model weight near one means that the model's complexity is perfectly justified. On the other hand, the almost equal weights of all variants suggest each model's justifiability is highly uncertain. The degree of complexity that is most justifiable is the complexity of the most complicated model within the set.

According to Schöniger [2016], a model justifiability analysis based on a model confusion matrix reveals the degrees of similarity between the alternative models. This insight is of importance in a multi-model setting and helps a modeler to reconsider a specific choice of prior model weights for models that some degree of similarity has been detected. Moreover, the confusion matrix makes the interpretation of the posterior model weights resulting from a conventional BMA/BMS easier. Using the latter, interpreting similar weights of two or more models is typically difficult. One conclusion is that the models are actually fairly similar in their predictions, while the similar model weights may just result from a similar overall goodness-of-fit. One can investigate the former case further via a model justifiability analysis and the resulting confusion matrix.

Practical implementation To compute the confusion matrix, one needs to compute the Bayesian model weights for all models in the set, adopting Equation (2.23). However, in the justifiability analysis, each competing model generates a finite series of prior predictions that serve as realizations of the "synthetic truth" instead of the measurement data $\mathcal{Y}$. These prior predictions are obtained by running models for N_{MC} samples drawn from the prior parameter distributions. Then, each synthetic data set is compared to the competing models by first computing the likelihood function as described in Equation (2.6), for example, for a single realization i of model $\mathcal{M}_k$ using the j-th data set of the model $\mathcal{M}_l$. The marginal likelihood (or BME) of the model $\mathcal{M}_k$ given the i-th realization of the model $\mathcal{M}_l$ can be obtained by calculating the arithmetic average of all likelihoods $p(\mathcal{M}_{l,j} \mid \mathcal{M}_k)$.

Figure 2.1 shows a schematic illustration of constructing a confusion matrix for two competing models $\mathcal{M}_k$ and $\mathcal{M}_l$. The green cell in this figure represents the likelihood

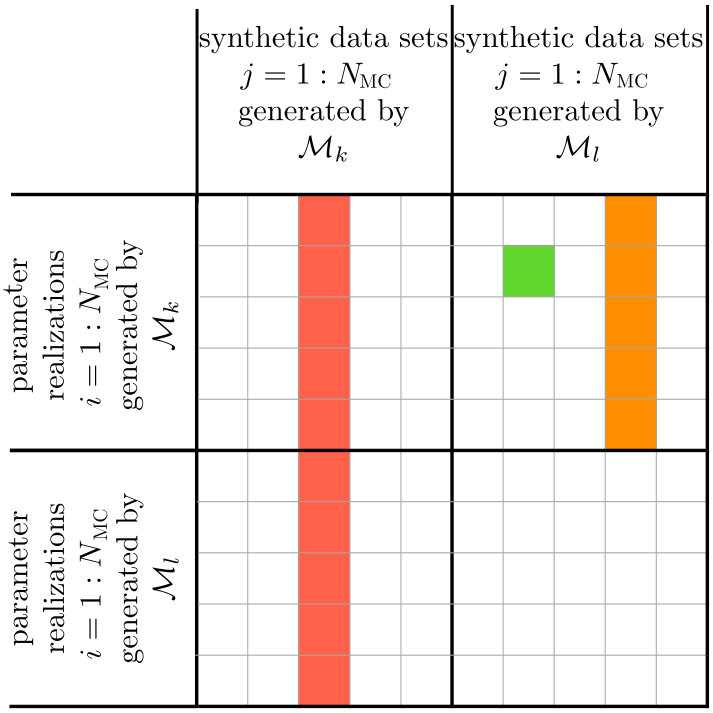


Figure 2.1: A schematic illustration of constructing the model confusion matrix

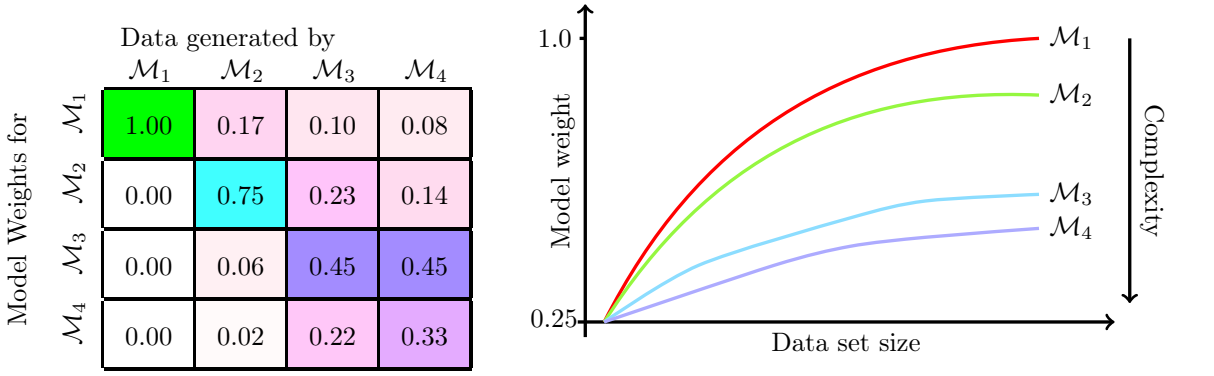
of a single realization of model $\mathcal{M}_k$ given a single realization of the reference model $\mathcal{M}_l$. The cell marked in orange shows the BME (averaged likelihood) of the model $\mathcal{M}_k$ given a single realization of the reference model $\mathcal{M}_l$. This BME value is then normalized by the sum of the BME values of all models given a single realization of the synthetic truth, shown in red, resulting in posterior model weights over all synthetic data of the reference model $\mathcal{M}_k$. In the last step, the values in each box, marked with bold borders, will provide the expected posterior weight ($\mathcal{W}^{post}$) of all models given that model $\mathcal{M}_k$ is true. This expected value can be written as

$$\begin{aligned} \mathcal{W}_{l|k}^{post} &= \frac{1}{N_{\mathrm{MC}}} \sum_{j=1}^{N_{\mathrm{MC}}} p\left(\mathcal{M}_l \mid \mathcal{M}_{k,j}\right) \\ &= \frac{1}{N_{\mathrm{MC}}^2} \sum_{j=1}^{N_{\mathrm{MC}}} \sum_{i=1}^{N_{\mathrm{MC}}} p\left(\mathcal{M}_{l,i} \mid \mathcal{M}_{k,j}\right). \end{aligned} \tag{2.25}$$

Interpretation of model justifiability results Figure 2.2a presents a didactic example of a model confusion matrix using the whole data set. The main diagonal entries reflect how well each model identifies itself as the data-generating process, given a specific

data set size. The values of the diagonal entries tend to be 1.00 using a data set with infinite size. However, for finite data sets, models might "confuse" their own predictions (misclassification) with that of the competing models. The off-diagonal entries of the model confusion matrix indicate the similarity between pairs of models. This finding can be helpful when comparing possible simplifications to a detailed reference model. A model confusion matrix helps to identify the model that provides the most identical results to the reference model at reduced computational cost [Schäfer Rodrigues Silva et al., 2020].

The schematic illustration in Figure 2.2b shows the model weights for the data-generating model as a function of the data set size and the complexity. The curves in the figure are based on the diagonal entries of the confusion matrix for increasing data set size. The



(a) Confusion matrix

(b) Model justifiability as a function of data set size and complexity

Figure 2.2: An example of a model justifiability analysis for a set of four models (adapted from Schöniger et al. [2015a]).

simplest model ($\mathcal{M}_1$) obtains a significantly higher weight than the competing models when it generated the data. This model weight reaches its maximum 100%, when all data are considered. This observation indicates that its complexity is perfectly supported by the largest data set. Following the same argumentation, the slightly more complex model $\mathcal{M}_2$ is also justified in all configurations, but with less confidence than the simpler model $\mathcal{M}_1$.

Model $\mathcal{M}_3$ can hardly be self-identified when using few observations. This means that either the complexity of this model is not justified given all the observations, and/or the models make very similar predictions such that they cannot be discriminated. When

including more data, the interpolated model can be self-identified and justified with increasing confidence. However, it never reaches a model weight of more than 50%, which means that there is never an "absolute majority" in favor of justifiability for this model.

The most complex model ($\mathcal{M}_4$) cannot be justified with these measurement configurations. Including more test data yields a clearer decision as indicated by the increasing deviation of the posterior model weights from the prior weights. However, this decision is in favor of the less complex, but similarly structured model $\mathcal{M}_3$, and not in favor of the model $\mathcal{M}_4$, even though it is now the one that in fact generated the data. Therefore, one can conclude that using all data points does not yet suffice to justify this model's high level of complexity and a drastically larger data set would be required to justify this model.

The justifiability method introduced above suggests which level of complexity could be derived from the chosen experimental setup. It does not aim at providing a ranking based on model adequacy. The combination of the latter with the justifiability analysis can be of great help in the task of model comparison.

Theoretical Upper Limit for Model Performance

Bayes Factor, in the context of Bayesian hypothesis testing, provides a performance comparison of pairwise competing models. However, in a validation benchmark task, we are also interested in comparing their performance to the best achievable performance. Schöniger et al. [2015b] argue that this theoretical upper limit for model performance exists when the measurement data set has noise. They propose that this limit can be established via determining a distribution of BME for a so-called *theoretically optimal model* (TOM), which is also dubbed as a sure-thing hypothesis by Jaynes [2003]. They define the observed data set as TOM, as it gives an exact fit with zero bias while having a minimum number of parameters, i.e., precisely zero, which is equivalent to zero variance. Stated differently, the TOM indicates the expected best possible performance in the presence of measurement error.

This theoretically optimal distribution of BME can be defined as the distribution of likelihoods of the observed data set given the perturbed data sets. Assuming that

measurement errors follow a Gaussian distribution and are independent and identically distributed, the TOM performance (shown as log-BME) has a distribution of the weighted sum of normal squared residuals. Consequently, this distribution can be defined by the chi-square distribution [Hald, 1998] as:

$$\chi^2(x) = \frac{1}{2^{(k/2)}\Gamma(k/2)} x^{k/2-1} \exp\left(-x/2\right), \tag{2.26}$$

with k being the number of degrees of freedom, which is equal to the size of the observed data set N. The upper limit of performance as represented by the TOM does not depend on the actual level of measurement error variance because the chi-square distribution is only a function of the data set size. Following Schöniger et al. [2015b], I do not include the TOM in the actual model ranking in this dissertation but use it only as an upper limit to the BME scale.

2.3 Two-stage surrogate-based Bayesian multi-model comparison

In this dissertation, the proposed framework includes a two-stage multi-model comparison based on the Bayesian perspective as suggested by Schöniger et al. [2015a]. First, a standard BHT and/or BMS is performed to rank the models based on the actually observed experimental data. This step is followed by a justifiability analysis in a synthetic setup, where model-generated data is used instead of the actual observed data from the experiment. The proposed two-step procedure separates model adequacy from model justifiability. The first stage suggests how the models are ranked in light of the actual experimental results, while the second stage helps to decide whether the most appropriate model from the conventional BMS analysis is really the best model in the model set or whether this model is only optimal given the limited amount of available measurement data for the chosen experimental setup. Additionally, the justifiability analysis provides insights about similarities among the tested models [Schöniger et al., 2015a].

To execute the two-stage Bayesian multi-model comparison explained above, one can add the measurement data to the model set under investigation. That means we add

the measurement data as a new row and column to the confusion matrix in Figure 2.1. This modification is shown in Figure 2.3. The cells marked in blue in this figure

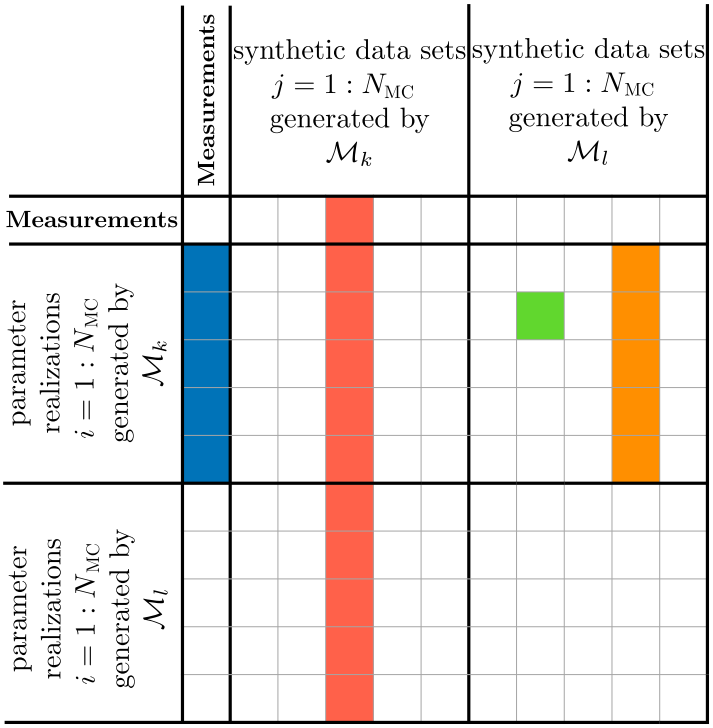


Figure 2.3: A schematic illustration of constructing the model confusion matrix for two-stage model comparison

represent a standard BMS procedure (Section 2.2.2) where the model $\mathcal{M}_k$ has been tested against the measurement data. These values can be computed by Equation (2.24) approximated via Monte Carlo integration techniques. The resulting extended model confusion matrix consists only of these entries, i.e., the bold boxes, and therefore has the size $(N_m + 1) \times (N_m + 1)$ for a set with N_m competing models and the measurement data.

The process described so far requires repetitive evaluations of the computational models. For computationally expensive models, e.g., those with run times in the order of minutes or hours, the Bayesian multi-model comparison becomes exceptionally intractable. One option to circumvent this challenge is to replace the original computational models with an easy-to-evaluate surrogate or proxy. The benefit of this substitution is two folds. On one hand, the overall required time for performing the model validation and comparison will drastically decrease. On the other hand, one

can take advantage of the surrogate model properties to scrutinize models further via sensitivity analysis, uncertainty quantification, and model calibration.

Surrogate models are, however, only approximations of full-complexity models. Consequently, any conclusions derived from BME values based on surrogates are only valid to the extent we are confident in the accuracy of the approximation. There are two possibilities to consider this potential inaccuracy in calculating the marginal likelihoods (BME), which is the basis for the model comparisons. First, one can estimate a surrogate-approximation error by comparing the prediction of the original and the surrogate models using a validation (test) set. This set must be different from the training set. A root-mean-square error can be used as a metric of approximation error. This value then can be included in the variance matrix $\mathbf{\Sigma}$ in Equation (2.6).

Nevertheless, it is not always feasible to perform so many simulations to split them into training and test sets. A notable example is when working with computationally demanding models. In these cases, care must be taken when interpreting model ranking/comparison results based on Bayesian model evidence for surrogate models. Using BME, a model may have been penalized for showing a low approximation quality; hence ranking a model via reduced models might be considered a conservative estimate of the accuracy of the full-complexity models. However, the opposite may also happen: the surrogate model might match the observed data better than the full-complexity model, just by chance. This would result in an overestimation of the evidence for the model. Therefore, the model ranking could be different for surrogate models than for full-complexity models, leading to wrong conclusions and possibly incorrect management decisions. Mohammadi et al. [2018] propose accounting for approximation errors explicitly by introducing a correction factor into model ranking to prevent such misleading results. The goal is to achieve model ranking results for full-complexity models using their surrogates that are more representative of the desired results. In what follows, the derivation of such a correction factor is presented.

As discussed earlier, the models are compared against the realizations of their counterparts in a justifiability analysis in the context of the Bayesian multi-model comparison. For computationally expensive models, the analysis must be performed using the surrogates for both the model under investigation and the reference model. BME values based on comparing surrogates against surrogates might also yield misleading comparison results. Therefore, the rest of this section is dedicated to derivation of a new

correction factor for a justifiability analysis. This extension is one contribution of this work to this research field.

2.3.1 Corrected BME for surrogate-based BMS

A robust surrogate-assisted Bayesian model selection requires corrected BME values. Let $\tilde{\mathcal{M}}_k$ be the surrogate representation of the full-complexity model $\mathcal{M}_k$ with an approximation error E_k so that $\tilde{\mathcal{M}}_k = \mathcal{M}_k + E_k$. Employing this property of additivity, Equation (2.24) can be reformulated as

$$p(\mathcal{Y} \mid \mathcal{M}_k) = \int_{\Theta_k} p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k + E_k, \boldsymbol{\theta}_k) P(\boldsymbol{\theta}_k \mid \mathcal{M}_k) d\boldsymbol{\theta}_k. \tag{2.27}$$

Assuming that the approximation errors are independent of the measurement errors in the definition of the likelihood function in Equation (2.6), Equation (2.27) can be rewritten as

$$p(\mathcal{Y} \mid \mathcal{M}_k) = \int_{\Theta_k} p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_k) p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_k) P(\boldsymbol{\theta}_k \mid \mathcal{M}_k) d\boldsymbol{\theta}_k, \tag{2.28}$$

where $p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_k)$ denotes the likelihood of the predictions generated by the original computational model Y_k given the approximate predictions $\tilde{Y}_k$ generated by its surrogate. This term can be computed by

$$p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_k) = \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{S}}} \exp\left(-\frac{1}{2} (Y_k(\boldsymbol{\theta}) - \tilde{Y}_k(\boldsymbol{\theta}))^T \boldsymbol{S}^{-1} (Y_k(\boldsymbol{\theta}_k) - \tilde{Y}_k(\boldsymbol{\theta}_k)) \right), \tag{2.29}$$

where $Y_k(\boldsymbol{\theta}_k)$ is the simulation output generated by the computational model for the parameter set $\boldsymbol{\theta}_k$. $\tilde{Y}_k(\boldsymbol{\theta}_k)$ is the corresponding surrogate approximation and $\boldsymbol{S}$ stands for the covariance matrix of approximation errors, which reflects the uncertainty in the approximation.

By multiplying and dividing the right-hand side of Equation (2.28) by the BME value of the surrogate model $p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k)$, we have

$$p(\mathcal{Y} \mid \mathcal{M}_k) = p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k) \int_{\Theta_k} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_k) \frac{p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_k) P(\boldsymbol{\theta}_k \mid \mathcal{M}_k)}{p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k)} d\boldsymbol{\theta}_k, \tag{2.30}$$

Using the Bayes theorem in Equation (2.4) and knowing that the parameter priors of the original and its surrogate model are equal, the BME of the computational model can be recast as

$$p(\mathcal{Y} \mid \mathcal{M}_k) = p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k) \int_{\Theta_k} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k \mid \tilde{\mathcal{M}}_k, \mathcal{Y}) d\boldsymbol{\theta}_k. \tag{2.31}$$

This representation of BME indicates that the corrected BME value of the computational model ($\mathrm{BME_{OM}}$) is a product of the BME value obtained by the surrogate model ($\mathrm{BME_{SM}}$) and the correction factor ($\mathrm{Weight_{SM}}$). Equation (2.31) can be written as

$$\mathrm{BME_{OM}} = \mathrm{BME_{SM}} \cdot \mathrm{Weight_{SM}}. \tag{2.32}$$

where

$$\begin{aligned} \mathrm{BME_{OM}} &= p(\mathcal{Y} \mid \mathcal{M}_k) \\ \mathrm{BME_{SM}} &= p(\mathcal{Y} \mid \tilde{\mathcal{M}}_k) \\ \mathrm{Weight_{SM}} &= \int_{\Theta_k} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k \mid \tilde{\mathcal{M}}_k, \mathcal{Y}) d\boldsymbol{\theta}_k. \end{aligned} \tag{2.33}$$

The correction factor $\mathrm{Weight_{surr}}$ is defined as the integral of the likelihood $p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_k)$ over the posterior parameter space. When using computationally expensive models, computing the likelihood for many samples from the posterior parameter space is not tractable. One strategy to approximate this correction factor is to reuse the simulations and predictions of the original and surrogate model in the training step, using

$$\mathrm{Weight_{SM}} \approx \sum_{i=1}^{P} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}^{\star}_{k,i}) \cdot p(\boldsymbol{\theta}^{\star}_{k,i} \mid \tilde{\mathcal{M}}_k, \mathcal{Y}). \tag{2.34}$$

This approach yields a corrected BME value that is more representative (although still an approximation) of the full-complexity model's BME. Moreover, the correction factor in Equation (2.33) could be approximated using only a maximum *a posteriori* parameter set $\boldsymbol{\theta}_k^{\mathrm{MAP}}$ as

$$\mathrm{Weight_{SM}} \approx \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{S}}} \exp\left(-\frac{1}{2} \boldsymbol{F}_{\star}^T \boldsymbol{S}^{-1} \boldsymbol{F}_{\star}\right), \tag{2.35}$$

where $\boldsymbol{F}_{\star}$ corresponds to the vector of approximation errors obtained for the maximum

a posteriori parameter set $\boldsymbol{\theta}_k^{\mathrm{MAP}}$, i.e., $\boldsymbol{F}_\star = \mathcal{M}_k(\boldsymbol{\theta}_k^{\mathrm{MAP}}) - \tilde{\mathcal{M}}_k(\boldsymbol{\theta}_k^{\mathrm{MAP}})$. This method also gives a very rough estimation of the correction factor and could be reliable only if the posterior was (very close to) a Dirac function.

2.3.2 Corrected model weights for surrogate-based justifiability analysis

In this section, the methodologies presented earlier in Section 2.2.2 are extended towards a surrogate-based Bayesian model justifiability analysis, where models are mutually examined against each other's predictions. Let us assume a case with two models, model $\mathcal{M}_k$ and model $\mathcal{M}_l$. The comparison of two models implies that one model, $\mathcal{M}_k$ in this case, is assumed to be the data generating process. The BME of the model $\mathcal{M}_l$ given the data generated by $\mathcal{M}_k$ takes the following form:

$$p(\mathcal{M}_k \mid \mathcal{M}_l) = \int_{\Theta_l} p(\mathcal{M}_k \mid \mathcal{M}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{M}_l) d\boldsymbol{\theta}. \tag{2.36}$$

For analyzing the computationally expensive models, one can calculate the BME value $p(\tilde{\mathcal{M}}_l \mid \tilde{\mathcal{M}}_k)$ of the easy-to-evaluate surrogate models, instead of computing the BME value for the original models $p(\mathcal{M}_l \mid \mathcal{M}_k)$. The surrogate representations of each analyzed model contains an approximation error: $\mathcal{M}_l = \tilde{\mathcal{M}}_l + \mathrm{E}_l$ and $\mathcal{M}_k = \tilde{\mathcal{M}}_k + \mathrm{E}_k$. Therefore, Equation (2.36) can be rewritten as:

$$p(\mathcal{M}_k \mid \mathcal{M}_l) = \int_{\Theta_l} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l + \mathrm{E}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{M}_l) d\boldsymbol{\theta} \tag{2.37}$$

$$p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l + \mathrm{E}_l, \boldsymbol{\theta}) = p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\mathcal{M}_l \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}). \tag{2.38}$$

By plugging (2.38) in (2.37), we get:

$$p(\mathcal{M}_k \mid \mathcal{M}_l) = \int_{\Theta_l} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\mathcal{M}_l \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{M}_l) d\boldsymbol{\theta}. \tag{2.39}$$

By multiplying and dividing the right-hand side of (2.39) by $p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l)$, the equation

can be written as:

$$p(\mathcal{M}_k \mid \mathcal{M}_l) = p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l) \int_{\Theta_l} \frac{p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{M}_l)}{p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l)} p(\mathcal{M}_l \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{2.40}$$

Knowing that $\frac{p(\mathcal{M}_k|\tilde{\mathcal{M}}_l,\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{M}_l)}{p(\mathcal{M}_k|\tilde{\mathcal{M}}_l)}$ is equal to $p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l, \mathcal{M}_k)$, one obtains:

$$p(\mathcal{M}_k \mid \mathcal{M}_l) = p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l) \int_{\Theta_l} p(\mathcal{M}_l \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l, \mathcal{M}_k) d\boldsymbol{\theta}. \tag{2.41}$$

The first term before integral can be cast as:

$$\begin{aligned} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l) &= \int_{\Theta_l} p(\tilde{\mathcal{M}}_k + \mathrm{E}_k \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l) d\boldsymbol{\theta} \\ &= \int_{\Theta_l} p(\tilde{\mathcal{M}}_k \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l) p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= p(\tilde{\mathcal{M}}_k \mid \tilde{\mathcal{M}}_l) \int_{\Theta_l} \frac{p(\tilde{\mathcal{M}}_k \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l)}{p(\tilde{\mathcal{M}}_k \mid \tilde{\mathcal{M}}_l)} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned} \tag{2.42}$$

Since $\frac{p(\tilde{\mathcal{M}}_k|\tilde{\mathcal{M}}_l,\boldsymbol{\theta})p(\boldsymbol{\theta}|\tilde{\mathcal{M}}_l)}{p(\tilde{\mathcal{M}}_k|\tilde{\mathcal{M}}_l)}$ is equal to $p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l, \tilde{\mathcal{M}}_k)$, we have:

$$p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_l) = p(\tilde{\mathcal{M}}_k \mid \tilde{\mathcal{M}}_l) \int_{\Theta_l} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l, \tilde{\mathcal{M}}_k) d\boldsymbol{\theta}. \tag{2.43}$$

By inserting (2.43) in (2.41), one obtains:

$$\begin{aligned} p(\mathcal{M}_k \mid \mathcal{M}_l) = p(\tilde{\mathcal{M}}_k \mid \tilde{\mathcal{M}}_l) &\times \int_{\Theta_l} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l, \tilde{\mathcal{M}}_k) d\boldsymbol{\theta} \\ &\times \int_{\Theta_l} p(\mathcal{M}_l \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l, \mathcal{M}_k) d\boldsymbol{\theta} \end{aligned} \tag{2.44}$$

or

$$\mathrm{BME}_{\mathrm{OMOM}} = \mathrm{BME}_{\mathrm{SMSM}} \cdot \mathrm{Weight}_{\mathrm{SM1}} \cdot \mathrm{Weight}_{\mathrm{SM2}}, \tag{2.45}$$

where

$$\begin{aligned}
\mathrm{BME}_{\mathrm{OMOM}} &= p(\mathcal{M}_k \mid \mathcal{M}_l) \\
\mathrm{BME}_{\mathrm{SMSM}} &= p(\tilde{\mathcal{M}}_k \mid \tilde{\mathcal{M}}_l) \\
\mathrm{Weight}_{\mathrm{SM1}} &= \int_{\Theta_l} p(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l, \tilde{\mathcal{M}}_k) d\boldsymbol{\theta} \\
\mathrm{Weight}_{\mathrm{SM2}} &= \int_{\Theta_l} p(\mathcal{M}_l \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\mathcal{M}}_l, \mathcal{M}_k) d\boldsymbol{\theta}.
\end{aligned} \tag{2.46}$$

Above, $\mathrm{BME}_{\mathrm{OMOM}}$ denotes the BME value when two original models are compared, while $\mathrm{BME}_{\mathrm{SMSM}}$ corresponds to the BME value when comparing two surrogate models. The latter can be computed by the method presented in Section 2.1.2, using the prediction of the model $\mathcal{M}_l$ evaluated on a certain model parameter set $\boldsymbol{\theta}$ instead of the measurement data $\mathcal{Y}$.

Similar to Section 2.3.1, one can use the training set of the surrogate models to compute the correction factors for both models:

$$\begin{aligned}
\mathrm{Weight}_{\mathrm{SM1}} &\approx \sum_{i=1}^{P} p\left(\mathcal{M}_k \mid \tilde{\mathcal{M}}_k, \boldsymbol{\theta}_i^*\right) p\left(\boldsymbol{\theta}_i^* \mid \tilde{\mathcal{M}}_l, \tilde{\mathcal{M}}_k\right) \\
\mathrm{Weight}_{\mathrm{SM2}} &\approx \sum_{i=1}^{P} p\left(\mathcal{M}_l \mid \tilde{\mathcal{M}}_l, \boldsymbol{\theta}_i^*\right) p\left(\boldsymbol{\theta}_i^* \mid \tilde{\mathcal{M}}_l, \mathcal{M}_k\right).
\end{aligned} \tag{2.47}$$

To extend the Bayesian justifiablity analysis for the computationally demanding models, the posterior model weights in Equation (2.25) need to be corrected by two factors $\mathrm{Weight}_{\mathrm{SM1}}$ and $\mathrm{Weight}_{\mathrm{SM2}}$ computed by Equation (2.47):

$$\begin{aligned}
\mathcal{W}_{l|k}^{post} &= \frac{1}{N_{\mathrm{MC}}} \sum_{j=1}^{N_{\mathrm{MC}}} p\left(\mathcal{M}_l \mid \mathcal{M}_{k,j}\right) \\
&= \frac{1}{N_{\mathrm{MC}}} \sum_{j=1}^{N_{\mathrm{MC}}} p\left(\tilde{\mathcal{M}}_l \mid \tilde{\mathcal{M}}_{k,j}\right) \cdot \ \mathrm{Weight}_{\ \mathrm{SM1}} \cdot \ \mathrm{Weight}_{\ \mathrm{SM2}}.
\end{aligned} \tag{2.48}$$

In this chapter, I introduced Bayes' theorem and showed how it is used for inference problems, such as calibration. Then, the validation workflow used in this work has been presented. The last section of this chapter presented the framework with a two-stage surrogate-based Bayesian multi-model comparison. The first stage suggests how

the models are ranked in light of the actual experimental results, while the second stage helps to decide whether the most appropriate model from the conventional BMS analysis is really the best model in the model set or whether this model is only optimal given the limited amount of available measurement data for the chosen experimental setup. Additionally, the justifiability analysis provides insights about similarities among the tested models.

Figure 2.4 illustrates the flow diagram of the proposed framework, which describes the actual workflow.

In the next chapter, I will show how we can make use of surrogate modeling to offset the computing cost of the framework when analyzing models with high runtimes.

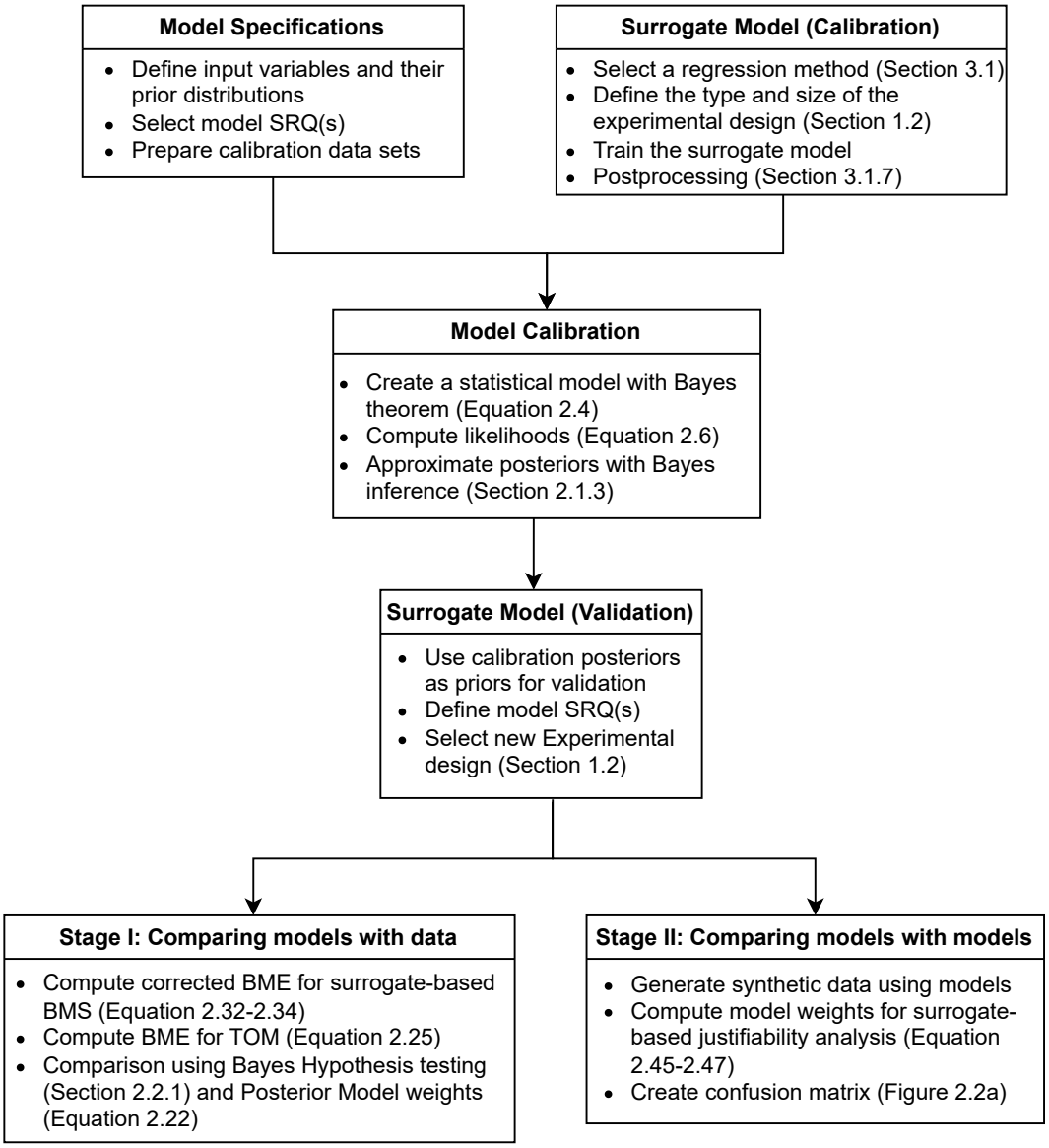


Figure 2.4: The workflow of the proposed validation framework.

3 Surrogate Modeling

The Bayesian framework, discussed in the previous chapter, requires the propagation of the parametric uncertainty through the given computational model. This task is also known as uncertainty propagation (UP). Typically, a significant number of model evaluations are required to yield statistical convergence. In practice, however, the computational complexity of the underlying computational model and the total available computational budget severely restrict the number of evaluations one can actually carry out. In such situations, the estimates produced by the Bayesian analysis lack sufficient trust, as the limited number of model evaluations can yield additional uncertainty.

The most common approach when dealing with computationally expensive models is to replace them with easy-to-evaluate surrogates. Simply put, one evaluates the model on a set of training (design) points and then strives to establish an accurate relationship between the model responses and the design points. Then, the original computational model can be substituted by its surrogate in the Bayesian analysis. Polynomial chaos expansion (PCE) is one of the most rigorous approaches to UP, thanks to its solid mathematical basis and ability to provide functional representations of stochastic quantities. However, the accuracy of the prediction of these surrogate models, trained with only a handful of simulations, is debatable. This argument is rooted in the fact that the surrogates do not attempt to quantify the epistemic uncertainty associated with their predictions.

This chapter highlights how a surrogate model using a PCE can be constructed for computationally intensive models with as few simulations as the computational budget allows. It is also shown how the Bayesian formalism can be materialized by employing the concept of PCE to account for the uncertainty in the surrogate's predictions. Moreover, we introduce a set of sequential adaptive sampling strategies in which one attempts to augment the initial design iteratively. In doing so, informative regions in

the parameter space are properly explored, avoiding the waste of computational resources as opposed to the so-called one-shot designs. These regions are more likely to provide valuable information on the behavior of the original model responses.

3.1 Polynomial Chaos Expansion

Over the past 20 years, the PCE methodology has become increasingly popular in the mathematical and engineering communities. In a probabilistic framework, uncertainties in input parameters are modeled via random variables. These input uncertainties can be investigated using PCE. This method provides the means to develop an approximation to the map between inputs and the SRQ(s). This mapping is both computationally tractable and sufficiently accurate. In addition to being a surrogate model, PCEs are also often used for uncertainty propagation and sensitivity analysis as one can analytically compute moments and sensitivity measures such as Sobol indices [Sudret, 2007].

The main idea of PCE is to expand an SRQ with a finite variance in a suitably built basis of multivariate polynomials that are orthogonal to the joint probability density functions of the inputs. It is worth noting that the random variables are assumed to be statistically independent or may be linearly correlated. The linear correlation can be handled by adequate linear transformation [Oladyshkin and Nowak, 2012]. In this thesis, I employ a non-intrusive regression-based PCE that works well for globally smooth problems, commonly used in many engineering applications. In contrast to intrusive PCE employed for solving stochastic PDEs and requiring code modifications, its non-intrusive variant treats the computational model as a black box.

A PCE is a linear regression that includes linear combinations of a fixed set of nonlinear functions with respect to the input variables, known as basis functions (Section 3.1.1). Let $\mathbf{X}$ be a d-dimensional random vector with independent components on a domain $\mathcal{D} \subset \mathbb{R}^d$. Moreover, consider the model $\mathcal{M} \in L^2_{f_\mathbf{X}}(\mathcal{D})$ with $L^2_{f_\mathbf{X}}(\mathcal{D})$ representing the space of all scalar-value models with finite variance in $f_\mathbf{X} = \prod_{i=1}^{d} f_{X_i}(x_i)$.

The PCE representation of the output random variable $\mathbf{Y}$ of the model $\mathcal{M}$ can be cast

as the following:

$$\mathbf{Y} = \mathcal{M}(\mathbf{X}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{X}), \tag{3.1}$$

where $\Psi_{\boldsymbol{\alpha}}(\mathbf{X})$ represents multivariate polynomials orthogonal with respect to $f_{\mathbf{X}}$ and $\boldsymbol{\alpha}$ denotes a multi-index that represents the components of the multivariate polynomials $\Psi_{\boldsymbol{\alpha}}$. The $c_{\boldsymbol{\alpha}} \in \mathbb{R}$ are the corresponding coefficients (coordinates). For practical reasons, the sum in Equation (3.1) needs to be truncated to a finite sum by introducing the truncated polynomial chaos expansion:

$$\mathcal{M}(\mathbf{X}) \approx \mathcal{M}^{PC}(\mathbf{X}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{X}), \tag{3.2}$$

where $\mathcal{A} \subset \mathbb{N}^d$ denotes the set of selected multi-indices of the multivariate polynomials. A standard truncation scheme can be defined as all polynomials in the d input variables of total degree less or equal to p:

$$\begin{aligned} \mathcal{A}^{d,p} &= \{\boldsymbol{\alpha} \in \mathbb{N}^d \ : \mid \boldsymbol{\alpha} \mid \leq p\} \\ \text{card } \mathcal{A}^{d,p} &\equiv P = \binom{d+p}{p}. \end{aligned} \tag{3.3}$$

A modification of the standard scheme, the hyperbolic truncation scheme uses the so-called q-norm to define the truncation [Blatman and Sudret, 2011], which reads

$$\begin{aligned} \mathcal{A}^{d,p,q} &= \left\{\boldsymbol{\alpha} \in \mathcal{A}^{d,p} : \|\boldsymbol{\alpha}\|_q \leq p\right\} \\ \|\boldsymbol{\alpha}\|_q &= \left(\sum_{i=1}^{d} \alpha_i^q\right)^{1/q} \end{aligned} \tag{3.4}$$

The hyperbolic truncation with $q < 1$ includes all the univariate high-degree terms but excludes high-degree terms with many interacting variables. For $q = 1$, hyperbolic truncation turns into the standard total-degree truncation scheme in Equation (3.3). For other truncation schemes, the reader is referred to Marelli et al. [2021].

3.1.1 Polynomial basis functions

The multivariate polynomials $\Psi_{\boldsymbol{\alpha}}(\mathbf{X})$ are the tensor product of the univariate polynomials:

$$\Psi_{\boldsymbol{\alpha}}(\mathbf{X}) := \prod_{i=1}^{d} \psi_{\alpha_i}^{(i)}(x_i). \tag{3.5}$$

The univariate orthonormal polynomials $\psi_{\alpha_i}^{(i)}(x_i)$ must satisfy the following expression:

$$\langle \psi_j^{(i)}, \psi_k^{(i)} \rangle := \int_{\mathcal{D}_{X_i}} \psi_j^{(i)}(x_i)\psi_k^{(i)}(x_i) f_{X_i}(x_i) dx_i = \delta_{jk}. \tag{3.6}$$

where i represents the input variable with respect to which they are orthogonal as well as the corresponding polynomial family, j and k the corresponding polynomial degree. $f_{X_i}(x_i)$ denotes the ith input marginal distribution and δ_{jk} is the Kronecker delta, which is equal to 1 when $j = k$ and 0 otherwise. The classical families of univariate orthonormal polynomials can be obtained by applying the Gram-Schmidt orthogonalization procedure [Strang and Freund, 1986] to the canonical family of monomials $\{1, x, x^2, \cdots\}$. The associated families of orthogonal polynomials for standard distributions are well-known. For example, for a uniformly distributed variable $X_i \sim \mathcal{U}(-1, 1)$, the family of polynomials is that of Legendre polynomials, whereas a normally distributed variable with zero mean and unit standard deviation $X_i \sim \mathcal{N}(0, 1)$ follow Hermite polynomials. Table 3.1 summarizes the univariate polynomial families of standard distributions [Sudret, 2007]. For detailed description of each of these classical families, the reader is referred to Xiu and Karniadakis [2002].

The calculation of polynomial basis via the classical families is grounded in the fact that exact knowledge about the probability density functions is available. However, the information about the parameter distribution is distinctly restricted in engineering applications, most importantly when environmental influences or natural phenomena are of interest or when prediction is involved. For instance, the material properties of subsurface reservoirs are not readily available to shed light on their distribution. This is also the case for the posterior parameter distribution obtained through a Bayesian calibration, which may not necessarily follow any known family of distributions presented in Table 3.1.

To overcome this challenge, Oladyshkin and Nowak [2012] demonstrate that statisti-

Table 3.1: Classical families of orthogonal polynomials

Type of variable	Distribution	Orthogonal polynomials	Hilbertian basis $\psi_k(x)$
Uniform $\mathcal{U}(-1,1)$	$1_{[-1,1]}(x)/2$	Legendre $P_k(x)$	$P_k(x)/\sqrt{\frac{1}{2k+1}}$
Gaussian $\mathcal{N}(0,1)$	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$	Hermite $H_{e_k}(x)$	$H_{e_k}(x)/\sqrt{k!}$
Gamma $\Gamma(a,\lambda=1)$	$x^a e^{-x}\mathbf{1}_{\mathbb{R}^+}(x)$	Laguerre $L_k^a(x)$	$L_k^a(x)/\sqrt{\frac{\Gamma(k+a+1)}{k!}}$
Beta $\mathcal{B}(a,b)$	$\mathbf{1}_{[-1,1]}(x)\frac{(1-x)^a(1+x)^b}{B(a)B(b)}$	Jacobi $J_k^{a,b}(x)$	$J_k^{a,b}(x)/\mathfrak{J}_{a,b,k}$
			$\mathfrak{J}^2_{a,b,k} = \frac{2^{a+b+1}}{2k+a+b+1}\cdot \frac{\Gamma(k+a+1)\Gamma(k+b+1)}{\Gamma(k+a+b+1)\Gamma(k+1)}$

cal moments are the only source of information that is propagated in all polynomial expansion-based stochastic approaches. The authors leverage this fact to propose a data-driven approach called arbitrary polynomial chaos expansion (aPCE), which can operate with probability measures that may be implicitly and incompletely defined via their statistical moments. Using aPCE, one can build the orthonormal polynomials even in the absence of the exact probability density function $f_{\mathbf{X}}(x)$. An aPCE generalizes chaos expansion techniques to arbitrary distributions with arbitrary probability measures, which can either be discrete, continuous or discretized continuous and can be specified either analytically (as probability density/cumulative distribution functions), numerically as a histogram, or as raw data sets. Unlike PCE, an aPCE only requires the existence of a finite number of moments at a finite expansion order and does not need complete knowledge or even the existence of a probability density function supported by limited available data. For more on the mathematical concept and the derivations, see the original paper [Oladyshkin and Nowak, 2012].

3.1.2 Calculation of the coefficients

Once the truncated multivariate basis $\Psi_{\boldsymbol{\alpha}}$ for the selected multi-index $\mathcal{A}$ in Equation (3.2) has been computed, the coefficients $c_{\boldsymbol{\alpha}}$ need to be estimated. There are two approaches for estimating the coefficients, namely intrusive and non-intrusive. Intrusive

computation schemes have been developed within the framework of stochastic finite element analysis [Ghanem and Spanos, 2003]. In an intrusive scheme, the constitutive equations of the physical problem are discretized both in the physical space via standard finite element techniques and in a random space via the polynomial chaos expansion. This procedure leads to coupled systems of equations that require ad-hoc solvers; thus, the term "intrusive" [Sudret, 2007, Debusschere, 2015].

On the other hand, non-intrusive methods make use of repeated runs, as in a Monte Carlo simulation. As a result, the computational model remains unchanged and is regarded as a black box that generates outputs for selected sets of random input vector $\mathbf{X}$. Non-intrusive methods make use of the multivariate orthogonal basis to compute the coefficients via the orthogonal projection of the random model response $\mathbf{Y}$ onto the corresponding basis function $\Psi_{\boldsymbol{\alpha}}(\mathbf{X})$, which can be written as

$$c_{\boldsymbol{\alpha}} = \mathbb{E}\left[\Psi_{\boldsymbol{\alpha}}(\mathbf{X}) \cdot \mathbf{Y}\right] = \int_{\mathcal{D}} \mathcal{M}(\boldsymbol{x})\Psi_{\alpha}(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}. \tag{3.7}$$

The integral in Equation (3.7) can be solved using Monte Carlo Simulation or numerical integration. The former estimates the expectation but has low efficiency. The latter uses Gaussian quadrature, sparse grids, or stochastic collocation methods to solve the integral on the right-hand side of Equation (3.7). For more detail, see the review paper by Xiu [2009].

The *least-square minimization* method, also known as regression approach, proposed in Berveiller et al. [2004, 2006], is regarded as a commonly used non-intrusive method. After selection of a truncation scheme $\mathcal{A}$, for example, $\mathcal{A}_{d,p}$ in (3.3), the infinite series in Equation (3.2) can be written as the sum of the truncated series and a residual ε:

$$\mathbf{Y} = \mathcal{M}(\mathbf{X}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\mathbf{X}) + \varepsilon. \tag{3.8}$$

The residual includes the error induced by truncated terms excluded from multi-index $\mathcal{A}$. The goal of the least-square minimization approach is to find the set of coefficients $\boldsymbol{c}$ that minimizes the mean square error, which reads

$$\mathbb{E}\left[\varepsilon^2\right] = \mathbb{E}\left[\left(\mathbf{Y} - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\mathbf{X})\right)^2\right]. \tag{3.9}$$

This minimization can be written as

$$\boldsymbol{c} = \arg \min_{c \in \mathbb{R}^{\mathrm{card}\,\mathcal{A}}} \mathbb{E}\left[\left(\mathcal{M}(\mathbf{X}) - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{X})\right)^2\right]. \tag{3.10}$$

To perform this minimization step, one can turn the problem to a discretized version of it in that the expectation operator in Equation (3.10) is substituted with the empirical mean over a training sample set

$$\hat{\boldsymbol{c}} = \arg \min_{c \in \mathbb{R}^{\mathrm{card}\,\mathcal{A}}} \frac{1}{N_{\mathrm{ED}}} \sum_{i=1}^{N_{\mathrm{ED}}} \left[\left(\mathcal{M}(\mathbf{x}^{(i)}) - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{x}^{(i)})\right)^2\right], \tag{3.11}$$

where $\mathbf{X} = \{\mathbf{x}^{(i)},\ i = 1, \cdots, N_{\mathrm{ED}}\}$ denotes the training sample set, also known as experimental design (ED) . This sample set is typically obtained by performing simulations using the input random vector $\mathbf{X}$ and will be discussed in detail in Section 3.2.

The minimization task in Equation (3.11) can be solved by taking the following steps:

1. The simulations with the computational model $\mathcal{M}$ are performed for each sample in ED and the resulting outputs are stored in a vector

$$\mathbf{Y} = \{\mathrm{Y}^{(1)} = \mathcal{M}(\mathbf{x}^{(1)}), \cdots, \mathrm{Y}^{(n)} = \mathcal{M}(\mathbf{x}^{(N_{\mathrm{ED}})})\}^\top. \tag{3.12}$$

2. A so-called *information matrix* is computed by evaluating the basis polynomials at each point in the ED:

$$\mathbf{A} = \left\{\mathbf{A}_{ij} \overset{\mathrm{def}}{=} \Psi_j\left(\mathbf{x}^{(i)}\right), i = 1, \ldots, N_{\mathrm{ED}}, \quad j = 1, \ldots, \mathrm{card}\,\mathcal{A}\right\}. \tag{3.13}$$

3. The solution to the minimization problem takes the following form:

$$\hat{\boldsymbol{c}} = \left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{Y}. \tag{3.14}$$

The surrogate model, which approximates the random response, denoted by $\hat{\mathbf{Y}}$, reads

$$\hat{\mathbf{Y}} = \mathcal{M}^{\mathrm{PCE}}(\mathbf{X}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \hat{\boldsymbol{c}}_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}. \tag{3.15}$$

Two issues require special attention in the minimization solution in Equation (3.14): ill-condition and well-posedness. As for the former, a singular value decomposition shall be employed to overcome the potential ill-conditioning of the information matrix [Sudret, 2014]. The latter requires that the number of unknowns $\operatorname{card} \mathcal{A}$ is smaller than the ED size $N_{\mathrm{ED}} = \operatorname{card} \mathbf{X}$. Sudret [2007] and Blatman [2009] suggest N_{ED} to be two or three times $\operatorname{card} \mathcal{A}$ as a rule of thumb.

This approach, however, becomes intractable for high dimensional input spaces or high-degree PCEs when considering computationally expensive models as the computational budget is limited. To overcome this, one can employ the concept of sparsity which offers elegant complexity control, over-fitting control, and feature extraction and has the potential for characterization of meaningful input variables along with the practical computational speed. As PCEs belong to linear regression models, employing the concept of sparsity can lead to zero values for many $c_{\boldsymbol{\alpha}}$ in the expansion in Equation (3.2), thus, fewer simulations are required for the surrogate training.

There are many mathematical approaches when dealing with a regression problem that leads to a sparse solution. These approaches have led to the emergence of numerous sparse solvers in the compressed sensing (e.g., Arjoune et al. [2017]) and in the sparse PCE. Lüthen et al. [2021] put the proposed solvers in the context of PCE into four categorizes: convex optimization solvers, greedy methods, iteratively re-weighted methods, and Bayesian sparse learning, also known as compressive sensing. For further details on different solvers in each category, the reader is referred to Lüthen et al. [2021].

In this chapter, I consider one greedy solver, Orthogonal Matching Pursuit, and two Bayesian sparse learning methods, FastARD and FastLaplace, whose mathematical concepts and algorithms follow. The Bayesian sparse learning methods can provide a probabilistic prediction, i.e., a prediction with the associated uncertainty. This prediction uncertainty can be used as the expected error when replacing the original computational model with a possibly less accurate surrogate. This property can be leveraged to refine the surrogate model and the Bayesian calibration and validation task.

Orthogonal matching pursuit

Greedy methods are stepwise regression methods where the regression terms (orthogonal polynomial basis) are added to the model one at a time according to a selection criterion. These methods aim to heuristically solve the intractable l^0-minimization problem. One of the most prominent greedy algorithms is Orthogonal Matching Pursuit (OMP) . It was first proposed by Pati et al. [1993] and explored extensively in the literature by e.g., Tropp and Gilbert [2007] and Doostan and Owhadi [2011]. In OMP, the regressors are iteratively retrieved one by one based on their correlation with the current residual approximation and are added to the active set of regressors. OMP uses OLS, described earlier in this section, to compute the PCE coefficients $\boldsymbol{c}$.

The OMP algorithm is a linear regression tool that minimizes the norm of the approximation residual at each iteration. The algorithm uses the leave-one-out error estimator (Section 3.1.3) to adaptively include the most relevant PCE basis term to the active set. Algorithm 1 summarizes the steps of the OMP solver.

Algorithm 1: Orthogonal Matching Pursuit

Inputs: Model evaluations $\mathbf{Y}$, regressors $\mathbf{\Psi}$

Initialize:

- $\mathbf{c}_{\boldsymbol{\alpha}}^{0} = 0, \quad \forall \boldsymbol{\alpha} \in \mathcal{A}^{d,p,q}$;
- Candidate set: $\Psi_{C,0} = \Psi_{\alpha}$;
- Active set: $\Psi_{A,0} = \varnothing$;
- Residual: $R_0 = \mathbf{Y}$

Result: Sparse regressor basis

1 **while** *itr number* $<$ *min*$(\operatorname{card} \mathcal{A}^{d,p,q}, N_{ED})$ *and* $\epsilon_{LOO} < \epsilon_{LOO}^{min}$ **do**
2 Find the polynomial $\Psi_{\boldsymbol{\alpha}_j}$ most correlated with the current approximation residual R_{j-1};
3 Add the polynomials $\Psi_{\boldsymbol{\alpha}_j}$ to the active set $\Psi_{A,j} = \Psi_{A,j-1} \cup \Psi_{\alpha_j}$;
4 Calculate the new polynomial coefficient $\mathbf{c}_{\boldsymbol{\alpha}}^{j}$ using $\Psi_{A,j}$ via OLS;
5 Calculate the new approximation residual via $R_j = \mathbf{Y} - \mathbf{\Psi}_{A,j}\mathbf{c}_{\alpha}^{j}$;
6 Calculate $\epsilon_{\mathrm{LOO}}^{j}$ via (3.38) and store it;
7 **end**
8 Select the active set of polynomials with the lowest ϵ_{LOO}

$\epsilon_{\mathrm{LOO}}^{min}$ in the stop criterion is the minimum value of ϵ_{LOO} for at least 10% of the maximum number of possible iterations. Moreover, $\operatorname{card} \mathcal{A}^{d,p,q}$ and N_{ED} denote the number of polynomial basis elements and the size of the experimental design.

Bayesian sparse learning solvers

Using Bayesian sparse learning, one imposes a sparsity-inducing prior on the coefficients (weights) of the predictors ($\Psi_{\boldsymbol{\alpha}}$ in the expansion (3.2)), whose parameters are considered to be random variables with a hyperprior. Then, the posterior of the weights are inferred via a sparse regression solver, such as the fast marginal likelihood maximization algorithm (FastARD) [Tipping et al., 2003] or its extension called FastLaplace [Babacan et al., 2009]. Henceforth, I will call the sparse extension of the aPCE, Bayesian sparse aPCE (BaSaPCE). This learning process leads to extremely sparse inferred predictors since they yield relatively few non-zero coefficients. That means a significant number

of the predictors give posterior distributions centered at zero and can be pruned out from the expansion in PCE.

BaSaPCE with FastARD Let the target variable be $\mathbf{Y}$, which is given by a surrogate function $\mathcal{M}^{\mathrm{PCE}}$ with an additive Gaussian noise that reads:

$$\mathbf{Y} = \mathcal{M}^{\mathrm{PCE}}(\mathbf{X}, \boldsymbol{c}) + \varepsilon, \tag{3.16}$$

where $\varepsilon \sim \mathcal{N}(0, \beta^{-1})$ is a zero mean Gaussian random variable with precision (inverse of variance) β. The uncertainty over the value of the target variable $\mathbf{Y}$ can be expressed by a probability distribution. Hence, the equation can be cast as:

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{c}, \beta) = \mathcal{N}(\mathbf{Y} \mid \mathcal{M}^{\mathrm{PCE}}(\mathrm{X}, \boldsymbol{c}), \beta^{-1}). \tag{3.17}$$

Let $\mathbf{X} = \{\mathrm{X}_1, \cdots, \mathrm{X}_{N_{\mathrm{ED}}}\}^\top$ be a data set of inputs with the corresponding model responses $\mathbf{Y} = \{\mathrm{Y}_1, \cdots, \mathrm{Y}_{N_{\mathrm{ED}}}\}^\top$. We can use the training set (ED) $\{\mathbf{X}, \mathbf{Y}\}$ to determine the values of the unknowns, $\boldsymbol{c}$ and β. Assuming these data points are drawn independently of the distribution in Equation (3.17), and using Equation (3.1), a multivariate Gaussian likelihood function can be derived as:

$$\begin{aligned} p\left(\mathbf{Y} \mid \mathbf{X}, \mathbf{c}, \beta\right) &= \prod_{i=1}^{N_{\mathrm{ED}}} \mathcal{N}(\mathrm{Y}_i \mid \boldsymbol{c}^\top \Psi(\mathrm{X}_i), \beta^{-1}) \\ &= \left(2\pi\beta^{-1}\right)^{-N_{\mathrm{ED}}/2} \exp\left(-\frac{\beta}{2}||\mathbf{Y} - \boldsymbol{c}^\top \Psi(\mathrm{X})||^2\right). \end{aligned} \tag{3.18}$$

We introduce a Gaussian prior distribution over the parameter vector $\boldsymbol{c}$ by giving each of the weight parameters c_i a separate hyper-parameter α_i. Thus, the prior of the weights reads as:

$$\begin{aligned} p(\boldsymbol{c} \mid \boldsymbol{\alpha}) &= \prod_{m=1}^{M} \mathcal{N}(c_m \mid 0, \alpha_m^{-1}) \\ &= \prod_{m=1}^{M} \left[(2\pi)^{-1/2} {\alpha_m}^{1/2} \exp\left(-\frac{1}{2}\alpha_m c_m^2\right)\right], \end{aligned} \tag{3.19}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_M\}^\top$ denotes the precision of the prior over its associated weight

parameter $\boldsymbol{c}$. M represents the number of the predictors, i.e., card $\mathcal{A}$. The form of prior is ultimately responsible for the sparsity properties of the model (for more details, see Tipping [2001]). Figure 3.1 displays the hierarchical framework of FastARD [Faul and Tipping, 2002, Tipping et al., 2003].

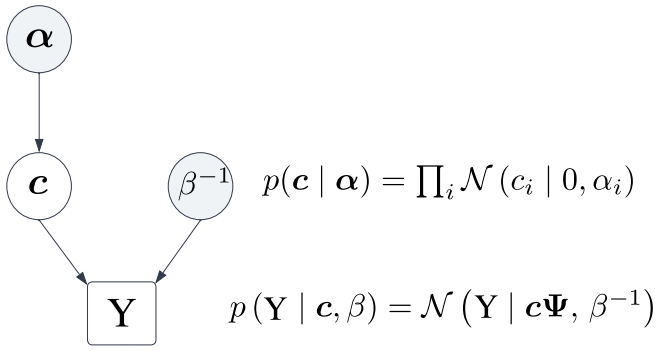


Figure 3.1: Graphical model of FastARD algorithm

The posterior distribution, conditioned on the model responses, is given by combining the likelihood in Equation (3.18) and the prior in Equation (3.19) according to Bayes' rule. This posterior, given $\boldsymbol{\alpha}$, can take the form:

$$p(\boldsymbol{c} \mid \mathbf{Y}, \boldsymbol{\alpha}, \beta) = \frac{p(\mathbf{Y} \mid \mathbf{X}, \mathbf{c}, \beta) p(\mathbf{c} \mid \boldsymbol{\alpha})}{p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\alpha}, \beta)}, \tag{3.20}$$

which also follows a Gaussian distribution, $\mathcal{N}(\mathbf{c} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with:

$$\begin{aligned} \boldsymbol{\mu} &= \beta \boldsymbol{\Sigma} \boldsymbol{\Psi}^{\top} \mathbf{Y} \\ \boldsymbol{\Sigma} &= \left(\mathcal{A} + \boldsymbol{\Psi}^{\top} \beta \boldsymbol{\Psi} \right)^{-1}, \end{aligned} \tag{3.21}$$

where $\boldsymbol{\Psi}$ is the design matrix of the size $N_{\mathrm{ED}} \times M$ with elements $\Psi_{ni} = \psi_i(x_n)$, and $\mathcal{A} = \mathrm{diag}(\alpha_i)$.

Marginal likelihood maximization: The values of $\boldsymbol{\alpha}$ and β can be determined via type-II maximum likelihood [Berger, 2013], also known as evidence approximation in the machine learning literature [Gull, 1989, MacKay, 1992a]. The ultimate goal in evidence approximation is to maximize the maximum likelihood function obtained by integrating

out the weight parameters $\boldsymbol{c}$:

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\alpha}, \beta) = \int p(\mathbf{Y} \mid \mathbf{X}, \mathbf{c}, \beta) p(\mathbf{c} \mid \boldsymbol{\alpha}) d\mathbf{c}. \tag{3.22}$$

This integral can be readily evaluated to obtain the log-likelihood form since it represents the convolution of two Gaussian distributions. The log-likelihood form reads as:

$$\begin{aligned} \ln p(\mathbf{Y} \mid \mathbf{X}, \alpha, \beta) &= \ln \mathcal{N}(\mathbf{Y} \mid 0, \mathbf{C}) \\ &= -\frac{1}{2} \left[N \ln (2\pi) + \ln |\mathbf{C}| + \mathbf{Y}^\top \mathbf{C}^{-1} \mathbf{Y} \right], \end{aligned} \tag{3.23}$$

where the matrix $\mathbf{C}$ with the size of $N_{\mathrm{ED}} \times N_{\mathrm{ED}}$ is defined as:

$$\mathbf{C} = \beta^{-1} \mathbf{I} + \boldsymbol{\Psi} \mathcal{A}^{-1} \boldsymbol{\Psi}^\top. \tag{3.24}$$

Let us denote the log marginal likelihood in Equation (3.23) by $\mathcal{L}(\boldsymbol{\alpha})$. Following Tipping et al. [2003], the Equation (3.23) can be written as:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}) &= -\frac{1}{2} \Bigg[N \ln (2\pi) + \ln |\mathbf{C}_{-i}| + \mathbf{Y}^\top \mathbf{C}_{-i}^{-1} \mathbf{Y} \\ &\qquad - \ln \alpha_i + \ln (\alpha_i + \Psi_i^\top \mathbf{C}_{-i}^{-1} \Psi_i) - \frac{(\Psi_i^\top \mathbf{C}_{-i}^{-1} \mathbf{Y})^2}{\alpha_i + \Psi_i^\top \mathbf{C}_{-i}^{-1} \Psi_i} \Bigg], \\ &= \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \frac{1}{2} \left[\ln \alpha_i - \ln (\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i} \right] \\ &= \mathcal{L}(\boldsymbol{\alpha}_{-i}) + l(\alpha_i), \end{aligned} \tag{3.25}$$

where $\mathbf{C}_{-i}$ is $\mathbf{C}$ with the contribution of basis vector i removed. The objective function has now been decomposed into $\mathcal{L}(\boldsymbol{\alpha}_{-i})$ and $l(\alpha_i)$. While the former is the marginal likelihood with Ψ_i excluded, the latter gives the same value for the isolated α_i. Moreover, the sparsity factor s_i and the quality factor q_i are defined as:

$$s_i := \Psi_i^\top \mathbf{C}_{-i}^{-1} \Psi_i \quad \text{and} \quad q_i := \Psi_i^\top \mathbf{C}_{-i}^{-1} \mathbf{Y}. \tag{3.26}$$

According to Tipping et al. [2003], the sparsity factor can be regarded as a measure of the extent that the basis factor Ψ_i *overlaps* those already present in the model. The

quality factor, however, can be seen as a measure of alignment of Ψ_i with the error of the model with that vector excluded. Analysis of $l(\alpha_i)$ in Faul and Tipping [2002] reveals that $\mathcal{L}(\boldsymbol{\alpha})$ has a unique maximum with respect to α_i as follows:

$$\alpha_i = \begin{cases} \frac{s_i^2}{q_i^2 - s_i}, & \text{if } q_i^2 > s_i, \\ \infty, & \text{otherwise.} \end{cases} \tag{3.27}$$

We compute q_i and s_i for all the basis functions Ψ_i in the pool $\boldsymbol{\Psi}$, including those not currently utilised in the model (i.e., for which $\alpha_i = \infty$). To do so, we iteratively maintain and update values s_i and q_i using the following expression:

$$s_m = \frac{\alpha_m S_m}{\alpha_m - S_m}, \qquad q_m = \frac{\alpha_m Q_m}{\alpha_m - S_m}, \tag{3.28}$$

where S_m and Q_m are obtained using:

$$S_m = \Psi_m^\top \beta \Psi_m - \Psi_m^\top \beta \boldsymbol{\Psi} \boldsymbol{\Sigma} \boldsymbol{\Psi}^\top \beta \Psi_m, \tag{3.29}$$

$$Q_m = \Psi_m^\top \beta \mathbf{Y} - \Psi_m^\top \beta \boldsymbol{\Psi} \boldsymbol{\Sigma} \boldsymbol{\Psi}^\top \beta \mathbf{Y}. \tag{3.30}$$

Here, quantities $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$ include only those basis functions that are currently included in the model. The sequential Bayesian sparse learning algorithm is presented in Algorithm 2. We start with an "empty" model, and sequentially include basis functions to increase the marginal likelihood, while modifying the weights. Further, we can increase the objective function within the same principled framework by removing basis functions which subsequently become redundant. For more details on the mathematical approach of the maximization, see Tipping et al. [2003].

It has been observed that the optimal values of many hyperparameters are typically infinite [Tipping, 2001]. Stated differently, parameter posterior distributions for many weights c_i are infinitely centered at zero. Consequently, the associated basis functions $\boldsymbol{\Psi}$ with these parameters play insignificant roles in the model's predictions and can be pruned out, leading to a sparse structure.

Prediction with BaSaPCE Having found values $\boldsymbol{\alpha}^*$ and β^* for the hyperparameters that maximize the marginal likelihood, one can evaluate the predictive distribution over

Algorithm 2: Bayesian sparse learning with FastARD algorithm

Inputs: Model evaluations $\mathbf{Y}$, regressors $\mathbf{\Psi}$, maximum iteration number N_{max}, and stopping threshold η

Initialize:

- $\beta \leftarrow$ initial value (e.g. $1/var[\mathbf{Y}]$);
- Select a single basis vector Ψ_i with the largest projection on targets $\mathbf{Y}$;
- Set $\alpha_i = \frac{||\Psi_i||^2}{||\Psi_i^\top \mathrm{Y}||^2/||\Psi_i||^2 - \beta^{-1}}$ and all other α_m to infinity ;

Result: The mean $\boldsymbol{\mu}$, the variance $\mathbf{\Sigma}$ of coefficients, and precision of the weight parameters $\boldsymbol{\alpha}$

for $n \leftarrow 1$ **to** N_{max} **do**
  Compute $\mathbf{\Sigma}$ and $\boldsymbol{\mu}$ of posterior distribution via (3.21);
  Calculate s_m and q_m for all M bases Ψ_m using Equations (3.27) and (3.29);
  Update β with $\beta = (N - M + \sum_m \alpha_m \mathbf{\Sigma}_{mm})/||\mathbf{Y} - y||^2$;
  Compute $\theta_i = q_i^2 - s_i$ of all M bases;
  if $\theta_i > 0$ *and* $\alpha_i < \infty$ **then**
    re-estimate α_i;
  else if $\theta_i > 0$ *and* $\alpha_i = \infty$ **then**
    add Ψ_i to the model with updated α_i;
  else
    delete Ψ_i from the model and set $\alpha_i = \infty$;
  end
  if *no features to add or delete* ***and*** *change of* $\boldsymbol{\alpha} < \eta$ **then**
    break;
  else
    Update precision parameters of coefficients $\alpha_i = s_i^2/\theta_i$;
  end
end

Y' for a new input $\mathbf{x}'$ by:

$$\begin{aligned} p(\mathrm{Y}' \mid \mathbf{x}', \mathbf{X}, \mathbf{Y}, \boldsymbol{\alpha}^*, \beta^*) &= \int p(\mathrm{Y} \mid \mathbf{x}', \mathbf{c}, \beta^*) p(\boldsymbol{c} \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\alpha}^*, \beta^*) d\mathbf{c} \\ &= \mathcal{N}(\mathrm{Y}' \mid \boldsymbol{\mu}^\top \mathbf{\Psi}(\mathbf{x}'), \sigma^2(\mathbf{x}')), \end{aligned} \tag{3.31}$$

where the variance of the predictive distribution is given by:

$$\sigma^2(\mathbf{x}') = (\beta^*)^{-1} + \Psi(\mathbf{x}')^\top \boldsymbol{\Sigma} \Psi(\mathbf{x}'). \tag{3.32}$$

$\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are calculated by Equation (3.21) in which $\boldsymbol{\alpha}$ and β are set to their optimized values $\boldsymbol{\alpha}^*$ and β^*. The first term in Equation (3.32) represents the noise on the data whereas the second term reflects the uncertainty associated with the coefficient $\boldsymbol{c}$.

BaSaPCE via FastLaplace Babacan et al. [2009] provide an extension of Relevance vector machine (RVM) [Tipping, 2001] called FastLaplace with an additional layer of hyperparameters. This approach has been proven to provide even sparser solutions compared to FastARD. Figure 3.2 illustrates the hierarchical framework used in FastLaplace. All quantities are random variables with distributions parametrized by hyperparameters. The hierarchical structure and the choice of priors and hyperpriors results in a sparsity-encouraging posterior distribution for the PCE coefficients $\boldsymbol{c}$. Each of the quantities and auxiliary variables is considered to be random with a

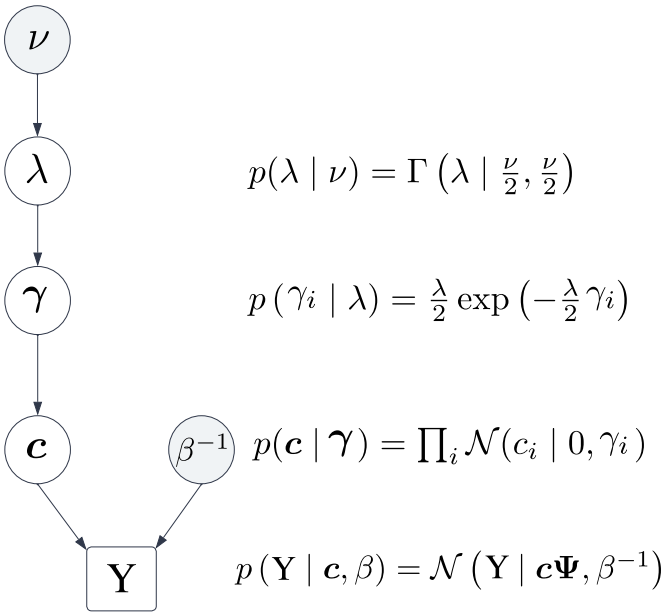


Figure 3.2: Graphical model of FastLaplace algorithm.

certain parametrized distribution, except for β, which is chosen *a priori* and $\nu = 0$. Figure 3.2 can be described as follows: The model outputs $\mathbf{Y}$ are assumed to be in-

dependent and identically distributed realizations of a random variable parameterized by PCE coefficients $\boldsymbol{c}$ and the parameter β (inverse of the noise variance parameter σ^2), which is a given hyperparameter of the framework. This results in the likelihood $p(\mathbf{Y} \mid \boldsymbol{c}, \beta) = \mathcal{N}(\mathbf{Y} \mid \boldsymbol{c}\boldsymbol{\Psi}, \beta^{-1})$.

The coefficients $\boldsymbol{c}$ are random variables, each following a normal distribution centered at zero and a variance of γ_i: $p(c_i \mid \alpha_i) = \mathcal{N}(c_i \mid 0, \gamma_i)$. The γ_i's are i.i.d and follow an exponential distribution with the shared parameter λ: $p(\gamma_i \mid \lambda) = \frac{\lambda}{2}\exp(-\frac{\lambda}{2}\gamma_i)$. The final layer in the framework, λ, is drawn from a Gamma distribution $p(\lambda \mid \nu) = \Gamma(\lambda \mid \frac{\nu}{2}, \frac{\nu}{2})$ with ν being zero, resulting in an uninformative prior. The goal is to compute the posterior distribution $p(\boldsymbol{c}, \boldsymbol{\gamma}, \lambda, \beta \mid \mathbf{Y})$ for a set of experimental design $\{\mathbf{X}, \mathbf{Y}\}$. Using Bayes' theorem, this posterior distribution can be computed by

$$p(\boldsymbol{c}, \boldsymbol{\gamma}, \lambda, \beta \mid \mathbf{Y}) = \frac{p(\boldsymbol{c} \mid \mathbf{Y}, \boldsymbol{\gamma}, \beta, \lambda) p(\boldsymbol{\gamma}, \beta, \lambda \mid \mathbf{Y})}{p(\mathbf{Y})} \tag{3.33}$$

with

$$p(\mathbf{Y}) = \iiiint p(\boldsymbol{c}, \boldsymbol{\gamma}, \lambda, \beta, \mathbf{Y}) d\boldsymbol{c}\, d\boldsymbol{\gamma}\, d\lambda\, d\beta. \tag{3.34}$$

For the sake of clarity, the dependency on ν is dropped. The marginal likelihood in (3.34) cannot be computed analytically. Similar to FastARD, one can perform the inference using type-II maximization likelihood [Berger, 2013]. Since the likelihood and prior (first and second terms in the denominator in Equation (3.33)) are normally distributed, the posterior distribution $\mathrm{p}(\boldsymbol{c}, \boldsymbol{\gamma}, \lambda, \beta \mid \mathbf{Y})$ also follows a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{c} \mid \boldsymbol{\mu}, \boldsymbol{\sigma})$, presented in Equation (3.21), with $\mathcal{A} = \mathrm{diag}(\gamma)$.

The logarithm of the marginal likelihood as a function of $\mathcal{L}(\boldsymbol{\gamma})$ takes the following form

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}) =& -\frac{1}{2}\left[\log|\mathbf{C}_{-i}| + \mathbf{Y}^\top \mathbf{C}_{-i}^{-1}\mathbf{Y} + \frac{\lambda}{2}\sum_{j\neq i}\gamma_j\right] \\ &+ \frac{1}{2}\left[\log\frac{1}{1+\gamma_i s_i} + \frac{q_i^2\gamma_i}{1+\gamma_i s_i} - \lambda\gamma_i\right] \\ =& \mathcal{L}(\boldsymbol{\gamma}_{-i}) + l(\gamma_i), \end{aligned} \tag{3.35}$$

where q_i and s_i are computed by Equation (3.26). A closed-form solution of the maximum of $\mathcal{L}(\boldsymbol{\gamma})$ is required to examine whether the i-th basis needs to be included in the model. This can be explored by taking the derivative of $\mathcal{L}(\boldsymbol{\gamma_i})$ or $l(\gamma_i)$ with respect to

γ_i and setting it to zero. The result is

$$\gamma_i = \begin{cases} \frac{-s_i(s_i+2\lambda)+s_i\sqrt{(s_i+2\lambda)^2-4\lambda\left(s_i-q_i^2+\lambda\right)}}{2\lambda s_i^2}, & \text{if } q_i^2 - s_i > \lambda \\ 0, & \text{otherwise.} \end{cases} \tag{3.36}$$

For further details on the derivation of the solution, see Babacan et al. [2009]. In the case of $\gamma_i = 0$, the associated basis/regressor ψ_i in the expansion has no contribution; thus, it is *inactive* and the corresponding coefficient c_i must obtain a zero value as well. On the contrary, a term with $\gamma_i > 0$ is considered to be *active* and included in the calculation. At every iteration, one regressor is selected. It can be either added to the pool of active basis, deleted from it, or reassessed via re-estimation of its variance γ_i. To do this, we compute the hypothetical updated value of γ_i for each regressor if this regressor alone was updated. Then, we calculate the associated hypothetical change in $\mathcal{L}(\boldsymbol{\gamma})$. To select the next basis vector to be added to the model, we only need to choose the basis Ψ that accounts for the largest increase in $\mathcal{L}$. This procedure is repeated until none of the regressors increases $\mathcal{L}$, or if the increase in $\mathcal{L}$ divided by the current value of $\mathcal{L}$ has been smaller than η twice in a row. Algorithm 3 summarizes the iterative algorithm of FastLaplace introduced by Babacan et al. [2009].

Algorithm 3: Bayesian sparse learning with FastLaplace algorithm

Inputs: Model evaluations $\mathbf{Y}$, regressors $\mathbf{\Psi}$, precision β, maximum iteration number N_{max}, and stopping threshold η

Initialize:

- Select the constant regression $\Psi_{\boldsymbol{\alpha_0}}$ as the only regressor in the model ;
- Set $\gamma_i = \frac{||\Psi_i||^2}{||\Psi_i^\top \mathbf{Y}||^2/||\Psi_i||^2 - \beta^{-1}}$ and all other γ_m to infinity ;
- Compute $\mathbf{\Sigma}$ and $\boldsymbol{\mu}$ of posterior distribution via (3.21);
- Calculate s_m and q_m for all M bases Ψ_m using Equations (3.27) and (3.29);
- Compute λ using $\lambda = \frac{N_{\text{ED}}-1}{\Sigma_{i=0}^{N_{\text{ED}}} \alpha_i/2}$;

Result: The mean $\boldsymbol{\mu}$, the variance $\mathbf{\Sigma}$ of coefficients, and precision of the weight parameters $\boldsymbol{\gamma}$

for $n \leftarrow 1$ **to** N_{max} **do**
  Select the basis ψ_i that accounts for the largest increase in $\mathcal{L}$;
  if *no features to add or delete* ***and*** *change of* $\mathcal{L}/\mathcal{L}_{current} < \eta$ **then**
    break;
  Compute $\delta = \text{nonzero}(\boldsymbol{\gamma}_{current}) - \text{nonzero}(\boldsymbol{\gamma})$;
  if $\delta = 0$ **then**
    re-estimate γ_i;
  else if $\delta = 1$ **then**
    add Ψ_i to the model with updated γ_i;
  else if $\delta = -1$ **then**
    delete Ψ_i from the model;
  Update $\mathbf{\Sigma}$ and $\boldsymbol{\mu}$;
  Update s_m and q_m;
  Update λ;
end

Babacan et al. [2009] use a fixed value for the hyperparameter $\beta = 100/\|\mathbf{Y}\|_2^2$. However, one can use k-fold cross-validation for computing β following Marelli et al. [2021].

Several candidates for βs are generated. The realizations $\mathbf{Y}$ are divided into k equally sized training and test chunks. Averaging over the k validation errors results in the corresponding cross-validation error. This process is repeated for each candidate β and the one with the smallest cross-validation error is selected. As the final step, the algorithm is run last time with the selected β and the entire experimental design to obtain the final solution.

3.1.3 On the accuracy of surrogate models

Once the surrogate model is constructed, its accuracy and predictive capability need to be assessed. In case, an independent set of inputs and outputs, also known as *validation set* $\{(\mathbf{x}_1^{Val}, \mathcal{M}(\mathbf{x}_1^{Val})), ..., (\mathbf{x}_{N_v}^{Val}, \mathcal{M}(\mathbf{x}_{N_v}^{Val}))\}$, is available in addition to a training set for training the surrogate model, the *validation error* can be computed as:

$$\varepsilon_{\text{Val}} = \frac{N_v - 1}{N_v} \left[\frac{\sum_{i=1}^{N_v} \left(\mathcal{M}(\mathbf{x}_i^{Val}) - \mathcal{M}^{PC}(\mathbf{x}_i^{Val}) \right)^2}{\sum_{i=1}^{N_v} \left(\mathcal{M}(\mathbf{x}_i^{Val}) - \hat{\mu}_{Y_{Val}} \right)^2} \right], \tag{3.37}$$

where $\hat{\mu}_{Y_{Val}} = \frac{1}{N_v} \sum_{i=1}^{N_v} \mathcal{M}(\mathbf{x}_i^{Val})$ denotes the mean of model responses for the validation set. Since computation of the aforementioned error requires a large number of model evaluations, it is only computationally tractable for models with low computational time.

To avoid additional model evaluations for assessing the accuracy of the surrogate model, an error based on the already evaluated ED is more desirable. One common approach is the leave-one-out cross validation (LOOCV) error, proposed by Geisser [1975], Stone [1974] explicitly introduced for PCE. This error, denoted by ε_{LOO}, is composed of rebuilding N_{ED} surrogate models in sequential ($\mathcal{M}^{PC\setminus i}$), using the original experimental design excluding i-th set ($\mathbf{X} \setminus \mathbf{x}^{(i)}$). Then the prediction error at the excluded set ($\mathbf{x}^{(i)}$) is computed. For more details, see Blatman and Sudret [2011].

Blatman [2009] shows that training independent surrogates is not needed, when using the linear superimposition of orthogonal terms, which is the case for PCE. Alternatively, the error can be calculated analytically from a single surrogate based on all sets in the

ED using the following equation:

$$\varepsilon_{\text{LOO}} = \frac{\sum_{i=1}^{N_{\text{ED}}} \left(\frac{\mathcal{M}(\mathbf{x}_i) - \mathcal{M}^{PC}(\mathbf{x}_i)}{1-h_i} \right)^2}{\sum_{i=1}^{N_{\text{ED}}} \left(\mathcal{M}(\mathbf{x}_i) - \mu_Y \right)^2}, \tag{3.38}$$

where h_i is the i-th diagonal entry of the experimental matrix $\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$, where $\mathbf{A}$ is the information matrix, defined in Equation (3.13). $\mu_Y = \frac{1}{N_{\text{ED}}} \sum_{i=1}^{N_v} \mathcal{M}(\mathbf{x}_i)$ is the mean of model responses for the training set.

3.1.4 Comparison of sparse solvers

In this section, I compare the following models: the Ishigami function, O'Hagan function and the borehole model. The Ishigami model is the widely-used highly nonlinear analytical function with three input parameters. The Ishigami function reads:

$$f\left(\theta_1, \theta_2, \theta_3\right) = \sin(\theta_1) + a \sin^2(\theta_2) + b\theta_3^4 \sin(\theta_1) \tag{3.39}$$

with uniform input $\boldsymbol{\theta} \sim \mathcal{U}([-\pi, \pi]^3)$. The typical choices for a and b are 7 and 0.1, respectively.

The borehole function computes the water flow through a borehole between two aquifers in m^3/yr [Harper and Gupta, 1983]. The function's simplicity and quick evaluation make it a popular choice in computer experiments for testing various uncertainty quantification methods. This function is also nonlinear and has an analytical form, but it is difficult to approximate. The borehole function takes the following form

$$f\left(r_w, L, K_w, T_u, T_l, H_u, H_l, r\right) = \frac{2\pi T_u \left(H_u - H_l\right)}{\ln\left(r/r_w\right)\left(1 + \frac{2LT_u}{\ln(r/r_w) r_w^2 K_w} + \frac{T_u}{T_l}\right)}. \tag{3.40}$$

Table 3.2 summarizes the input random variables and the respective distributions.

In Table 3.2, the normal distribution is defined with mean μ and variance σ^2. The log-normal distribution of the last variable is defined with parameters μ and σ^2 such that the natural logarithm of this variable follows a normal distribution.

Table 3.2: Borehole function: Input random variables and their distributions

Parameter name	Range/Parameters	Unit	Distribution type
Borehole radius, r_w	$(0.10, 0.0161812)$	m	normal
Borehole length , L	$[9855, 12045]$	m	uniform
Borehole hydraulic conductivity, K_w	$[9855, 12045]$	m/yr	uniform
Transmissivity of upper aquifer, T_u	$[63070, 115600]$	m^2/yr	uniform
Transmissivity of lower aquifer, T_l	$[63.1, 116]$	m^2/yr	uniform
Potentiometric head of upper aquifer, H_u	$[990, 1110]$	m	uniform
Potentiometric head of lower aquifer, H_u	$[700, 820]$	m	uniform
Radius of influence, r	$(7.71, 1.0056)$	m	log-normal

As the third model, I use the O'Hagan function [Oakley and O'Hagan, 2004], which reads

$$f(\boldsymbol{\theta}) = \mathbf{a_1}^\top \boldsymbol{\theta} + \mathbf{a_2}^\top \sin(\boldsymbol{\theta}) + \mathbf{a_3}^\top \cos(\boldsymbol{\theta}) + \boldsymbol{\theta}^\top M \boldsymbol{\theta}. \tag{3.41}$$

The values of the coefficients vectors $\mathbf{a_1}$, $\mathbf{a_2}$ and $\mathbf{a_3}$ and the Matrix M are known. The input random variables $\boldsymbol{\theta}$ are independent and follow normal distributions $\theta_i \sim \mathcal{N}(0, 1)$ for $i = 1, \cdots, 15$.

For each model, a fixed basis is used. Generally, the best total degree p and the hyperbolic truncation q are not *a priori* known. However, here for the sake of comparability, the following values are selected:

- Ishigami function ($d = 3$): p=14 and q=1 with 680 regressors
- Borehole function ($d = 8$): p=5 and q=1 with 1287 regressors.
- O'Hagan function ($d = 15$): p=7 and q=0.65 with 1401 regressors.

As for the experimental design, ten and thirty samples were selected via the LHS method as the initial experimental design for the Ishigami and Borehole functions, respectively. These initial experimental designs were then augmented iteratively with a space-filling approach to generate an experimental design in which the samples in the training set are distributed evenly over the design (parameter) space. The details are presented in Section 3.2.1. To account for the variabilities in the results, 50 replications of the computations were performed.

Figures 3.3 to 3.5 show boxplots of the validation error over the increasing size of the experimental design for three sparse learning solvers, namely OMP, BaSaPCE via

FastARD, and FastLaplace versus the commonly used OLS method. Unlike the benchmarking effort in Lüthen et al. [2021], I used aPCE instead of the regular PCE. The difference between these approaches is highlighted in Section 3.1.1. In the plots, the lines and the line inside the boxplots denote the median of the validation error.

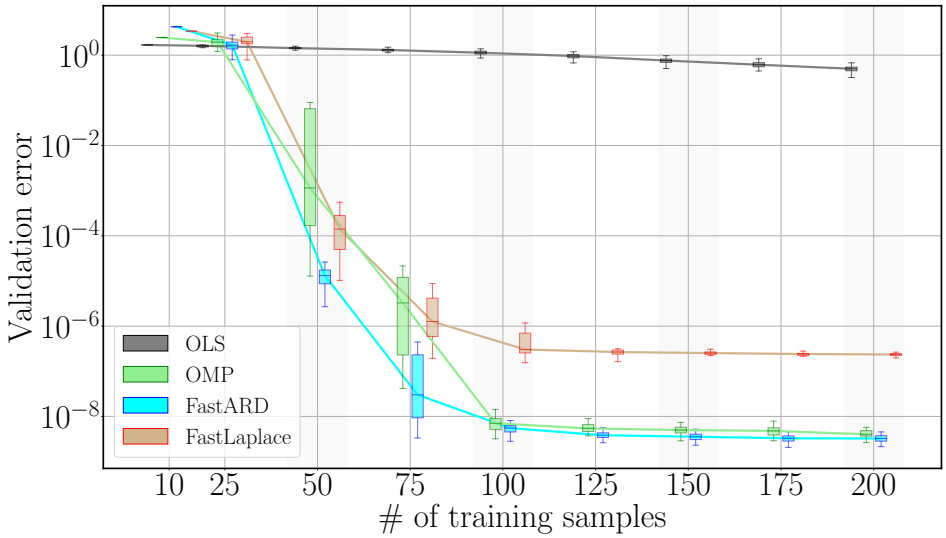


Figure 3.3: Results of solver comparison for the Ishigami model ($d = 3$, $p = 14$, $q = 1$)

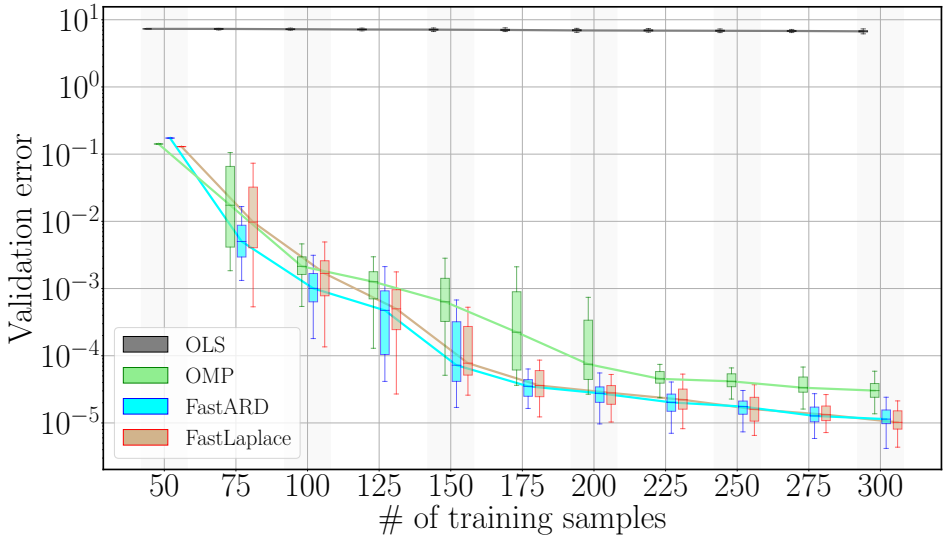


Figure 3.4: Results of solver comparison for the borehole function ($d = 8$, $p = 4$, $q = 1$).

The findings of the solver comparison can be summarized as follows:

- For the low-dimensional Ishigami function, all solvers perform similarly for small

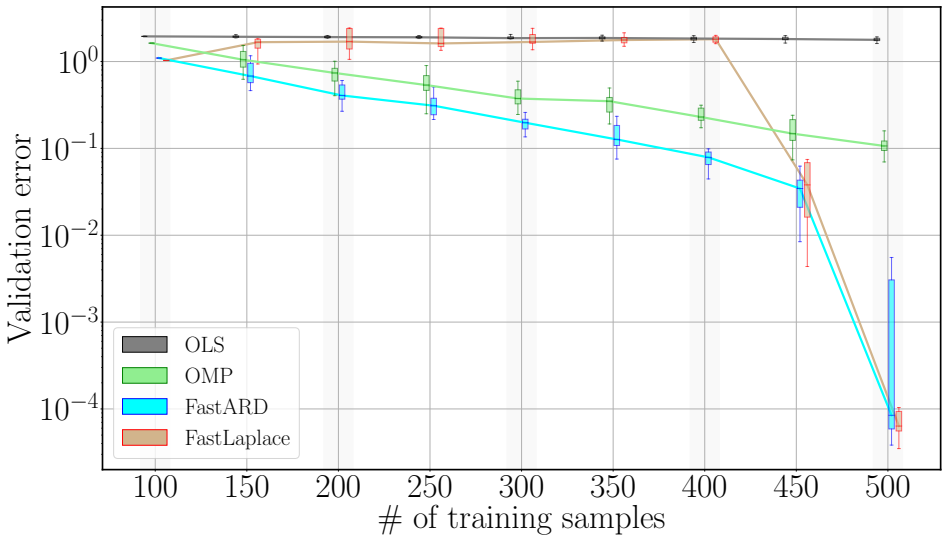


Figure 3.5: Results of solver comparison for the O'Hagan model ($d = 15$, $p = 7$, $q = 0.5$)

experimental designs. However, for larger experimental designs, a significant difference can be noticed.

- With increasing training samples, OMP and FastARD outperform FastLaplace for the Ishigami function.
- For the 8-dimensional borehole model, however, the Bayesian sparse learning solvers, namely FastARD and FastLaplace show lower validation errors, especially for larger experimental designs.
- For O'Hagan's model with 15 parameters, OMP can not show good accuracy and has a lower error reduction rate with increasing training samples compared to the Bayesian sparse learning methods.
- Comparing Bayesian sparse learning methods, FastARD and FastLaplace for the high dimensional example, it can be observed that the FastARD algorithm tends to have better accuracy (lower validation error) for the smaller size of the experimental design. However, as the experimental design gets larger, the FastLaplace algorithm outperforms FastARD.
- The OLS solver shows no significant change in validation error for all considered numerical examples with the increasing experimental design.

3.1.5 Treating spatial and temporal dependencies

What is presented so far is applicable for single-output computational models, i.e., the estimated outputs are only delivered for one instance in time or space. In many engineering applications, most of the developed computational models give responses at a fixed set of discrete time instances $\mathbf{T} = \{t_1, \cdots, t_{n_t}\}$ and spatial positions $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \{(\mathrm{x}_1, \mathrm{y}_1, \mathrm{z}_1), \cdots, (\mathrm{x}_{n_s}, \mathrm{y}_{n_s}, \mathrm{z}_{n_s})\}$ for many SRQs $\mathbf{R} = \{\mathbf{R}_1, \cdots, \mathbf{R}_{n_{out}}\}$. Following the commonly used approach of individual surrogate modeling that treats each of these scalar outputs separately, the training and prediction costs dramatically increase, which is proportional to $n_t \times n_s \times n_{out}$. Furthermore, individual surrogate modeling may lead to a high degree of redundancy, as the simulation outputs are highly correlated [Nagel et al., 2020].

To circumvent this computational issue, Higdon et al. [2008] proposed to treat the temporal dependency of the outputs by deploying dimensionality reduction techniques in a way that the correlations in space and time are preserved. Dimensionality reduction involves the compression of high-dimensional data into a representation with reduced dimensionality. Moreover, it is important in many fields, since it mitigates the curse of dimensionality and other unwanted characteristics of high-dimensional spaces.

Principal component analysis (PCA) is a well-established method for analyzing the relationships between dependent variables and obtaining linearly uncorrelated variables with decreasing variances by applying an orthogonal transformation. Recently, several attempts have been made to introduce PCA for uncertainty quantification. Aversano et al. [2019] employed PCA to compress a large number of predicted SRQs into a smaller set of scalars and trained surrogate models using a combination of PCA with Kriging for uncertainty analysis. Manfredi and Trinchero [2020] and Memon et al. [2020] used PCA for data compression and fed it to two compact surrogate models (i.e., sparse PCE and least-square support vector machine regression) for efficient uncertainty quantification. Nagel et al. [2020] applied PCA in conjunction with a sparse PCE to perform sensitivity analysis and calibration for a stormwater management model.

Let us consider a random vector $\mathbf{Y} \in \mathbb{R}^{n_t \times n_s \times n_{\mathrm{out}}}$ containing N realizations of model outputs $\mathbf{Y} = \{\mathrm{Y}^{(1)}, \cdots, \mathrm{Y}^{(N)}\}^\top$. The first two moments, mean ($\boldsymbol{\mu}_{\mathbf{Y}} = \mathbb{E}[\mathbf{Y}]$) and the

covariance ($\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathrm{Cov}\,[\mathbf{Y}]$) are approximated by

$$\bar{\boldsymbol{\mu}}_{\mathbf{Y}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{Y}^{(n)}, \quad \bar{\boldsymbol{\Sigma}}_{\mathbf{Y}} = \frac{1}{N-1}\sum_{n=1}^{N} \left(\mathrm{Y}^{(n)} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}}\right)\left(\mathrm{Y}^{(n)} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}}\right)^{\top}. \tag{3.42}$$

The PCA transformation can be performed by

$$\tilde{\boldsymbol{\zeta}}^{(n)} = \bar{\boldsymbol{\Phi}}_{N^\star}^{\top}\left(\mathrm{Y}^{(n)} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}}\right) \quad \text{for } n = 1, \cdots, N, \tag{3.43}$$

where $\bar{\boldsymbol{\Phi}}_{N^\star}^{\top} = (\bar{\Phi}_1, \cdots, \bar{\Phi}_{N^\star})$ are composed of $N^\star < N$ eigenvectors for which the proportion $\sum_{i=1}^{N^\star} \bar{\lambda}_i / \sum_{i=1}^{N} \bar{\lambda}_i$ of the total empirical variance is equal or larger than a user-defined threshold. The eigenvectors $\bar{\Phi}_i$ and the eigenvalues $\bar{\lambda}_i$ must meet the following condition

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{Y}}\bar{\Phi}_i = \bar{\lambda}_i\bar{\Phi}_i \quad \text{for } i = 1, \ldots, N. \tag{3.44}$$

where the eigenvalues are arranged in the descending order $\bar{\lambda}_0 \geq \bar{\lambda}_1 \geq \cdots \geq \bar{\lambda}_N$. Equation (3.43) represents the reduced PCA representation of Y_n in terms of the principal components $\tilde{\zeta}_i^{(n)} = \bar{\Phi}_i^{\top}\left(\mathrm{Y}^{(n)} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}}\right)$ for $i = 0, \cdots, N^\star$. The reduced model output can be represented by

$$\mathbf{Z} = \begin{pmatrix} \tilde{\boldsymbol{\zeta}}^{(1)\top} \\ \tilde{\boldsymbol{\zeta}}^{(2)\top} \\ \vdots \\ \tilde{\boldsymbol{\zeta}}^{(N)\top} \end{pmatrix} = \begin{pmatrix} \tilde{\zeta}_0^{(1)} & \tilde{\zeta}_1^{(1)} & \cdots & \tilde{\zeta}_{N^\star}^{(1)} \\ \tilde{\zeta}_0^{(2)} & \tilde{\zeta}_1^{(2)} & \cdots & \tilde{\zeta}_{N^\star}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\zeta}_0^{(N)} & \tilde{\zeta}_1^{(N)} & \cdots & \tilde{\zeta}_{N^\star}^{(N)} \end{pmatrix}. \tag{3.45}$$

The expression in Equation (3.45) shows the compressed data set that maintains the majority of the total variation. One can reconstruct the original observed samples $n = 1, \cdots, N$ approximately by

$$\mathrm{Y}^{(n)} \approx \bar{\boldsymbol{\mu}}_{\mathbf{Y}} + \bar{\boldsymbol{\Phi}}_{N^\star}^{\top}\tilde{\boldsymbol{\zeta}}^{(n)} = \bar{\boldsymbol{\mu}}_{\mathbf{Y}} + \sum_{i=0}^{N^\star} \tilde{\zeta}_i^{(n)}\bar{\Phi}_i. \tag{3.46}$$

Once the transformation of the model outputs $\mathbf{Y}$ are obtained by a few principal components, a surrogate model $\tilde{\zeta}_p(\mathbf{X}) \approx \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \hat{\boldsymbol{c}}_{p,\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\mathbf{X})$ for $p = 0, \cdots, N^\star$ via BaSaPCE (Section 8) can map the input parameter set $\mathbf{X}$ to the principal components $\mathbf{Z}$. For a new input $\mathbf{x}'$, Equations (3.31) and (3.32) yield the predictive distribution over $\tilde{\zeta}'$. The

conditional expectation of Y' with respect to $\mathbf{x}'$ can be approximated by

$$\mathbb{E}\left[\mathrm{Y}' \mid \mathbf{x}'\right] \approx \bar{\boldsymbol{\mu}}_{\mathbf{Y}} + \sum_{p=0}^{N^\star} \tilde{\zeta}_p(\mathbf{x}')\bar{\Phi}_p \approx \bar{\boldsymbol{\mu}}_{\mathbf{Y}} + \sum_{p=0}^{N^\star} \left(\sum_{\boldsymbol{\alpha}\in\mathcal{A}} \hat{\boldsymbol{c}}_{p,\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{x}') \right) \bar{\Phi}_p. \tag{3.47}$$

Note that when using more and more components, the metamodel error increases. According to Blatman and Sudret [2013], it is required to dynamically augment the experimental design's size dynamically to have better global control over the error. Doing so allows us to capture the details of the higher-order principal components.

3.1.6 Bootstrap-based prediction confidence interval

To assess the variability of the estimated coefficients $\boldsymbol{c}$ in Equation (3.16), one can make use of the *bootstrap* resampling method [Efron, 1982]. Let us assume that the SRQ, ω, is a function of a training set (experimental design) $\mathbf{X}$ with a finite size N_{ED}. For bootstrapping, it is required to construct b new sample sets from the original training set $\mathbf{X} = \{\mathcal{X}^{(1)}, \cdots \mathcal{X}^{(b)}\}$ and $\mathbf{Y} = \{\mathcal{Y}^{(1)}, \cdots \mathcal{Y}^{(b)}\}$. This resampling can be performed with substitution, which means b times N_{ED} realizations might include the same realization multiple times. Then, a family of b surrogate models is trained using the resampled training subsets. The estimated quantities from each of the b surrogate models, $\Omega = \{\omega^{(1)}, \cdots \omega^{(b)}\}$, yield a set of estimates, that can be directly used to assess the variability of the SRQ due to the finite size of the training set $\mathbf{X}$.

Marelli and Sudret [2018] proposed using the bootstrap technique to provide local error estimates (i.e., confidence interval) to the PCE predictions. This can be achieved by resampling the PCE coefficients using $\boldsymbol{c}_{\boldsymbol{\alpha}}$ via bootstrapping. To do so, a set of coefficients $\boldsymbol{c}_{\boldsymbol{\alpha}}^{(i)}$ is estimated for each generated design in the set of bootstrapped experimental design $\{\mathcal{X}^{(i)}, \mathcal{Y}^{(i)}\}$ with $i = 1, \cdots, b$. Consequently, the prediction of each PCE using Equation (3.15) at a new point $\boldsymbol{x}$ can be collected in a set of full responses (trajectories) $\{\tilde{\mathrm{Y}}^{(i)}(\boldsymbol{x}), i = 1, \cdots, b\}$. Then, one can employ the empirical quantities to provide a confidence interval on the PCE prediction at each new sample $\boldsymbol{x}$, or any other derived quantities, such as sensitivity indices [Dubreuil et al., 2014] or probability of failure in structural reliability applications [Notin et al., 2010, Picheny et al., 2010, Marelli and Sudret, 2018, Cheng and Lu, 2020, Guo et al., 2020]. The process of bootstrap based

resampling to better estimate point-wise confidence intervals has extensively been investigated in the Gaussian process modeling literature, see e.g., [Den Hertog et al., 2006, Van Beers and Kleijnen, 2008, Kleijnen and Van Beers, 2013, Mehdad and Kleijnen, 2015].

The computational complexity of the bootstrap-PCE is significant, especially for large experimental designs ($N \sim 1000$) and high dimensional input parameters when using BaSaPCE or any other sparse regression method. To circumvent this problem, Marelli and Sudret [2018] suggest using a *fast* version of bootstrap-PCE, in that the sparse polynomial bases are determined first by the sparse learning algorithm for the full experimental design $\mathbf{X}$. This step is followed by bootstrapping, which only recalculates the coefficients ($\boldsymbol{c_\alpha}$) for other bootstrap-resampled experimental designs via an OLS method on the identified sparse basis. The deployment of the classical OLS method can be justified by the fact that accurate PCE predictions are not necessary for estimating confidence intervals. However, Marelli and Sudret [2018] recommend employing *full* bootstrapping for computationally expensive models, whose single run may take several hours. For the *full* bootstrapping, the (Bayesian) sparse learning algorithm is used to compute the PCE coefficients for each of the bootstrap-resampled experimental designs.

As discussed earlier, BaSaPCE provides a prediction error for a new sample $\boldsymbol{x}$ using Equation (3.32). This prediction confidence bound reflects the variability due to the choice of polynomial terms included in the expansion. Using bootstrap-BaSaPCE, however, includes the variability due to the finite size of the experimental design. Moreover, when treating the spatial and temporal dependencies of the model outputs with dimensionality reduction methods, transforming the prediction uncertainty, estimated by Equation (3.32), computed for the principal components to the original output space is not trivial. Employing bootstrapping, in turn, allows us to produce trajectories in the reduced space and transform them back to the output space. Consequently, one can use these trajectories to compute the prediction confidence bounds.

3.1.7 Surrogate model properties

Once the PC bases have been set up, and the coefficients have been estimated, the post-processing of the expansion provides valuable metrics that help analyze the model under investigation. These metrics can reveal the average behavior of the SRQs and their

spreads using mean and variance, respectively. Moreover, one can estimate confidence intervals of SRQs and perform a sensitivity analysis, which shows how the variability of the model response quantities is affected by the variability of each input variable or combinations thereof.

Moment analysis The mean and standard deviation of a PCE truncated series $\tilde{\mathrm{Y}} = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \boldsymbol{c_\alpha}\Psi_{\boldsymbol{\alpha}}(\mathbf{X})$ can be readily computed given the orthonormality of the PCE basis expressed in Equation (3.6). Since each polynomial is orthogonal to $\Psi_{\mathbf{0}} \equiv 1$, that is $\mathbb{E}\left[\Psi_{\boldsymbol{\alpha}}(\mathbf{X})\right] = 0 \ \forall \boldsymbol{\alpha} \neq 0$. Therefore, the first term of the expansion represents the expected value of $\hat{\mathrm{Y}}$:

$$\mathbb{E}[\tilde{\mathrm{Y}}] = \mathbb{E}\left[\sum_{\alpha\in\mathcal{A}} \boldsymbol{c_\alpha}\Psi_{\alpha}(\mathbf{X})\right] = c_0. \tag{3.48}$$

Similar to the mean, the variance of $\hat{\mathrm{Y}}$ reads

$$\sigma^2_{\tilde{\mathrm{Y}}} = \mathrm{Var}[\tilde{\mathrm{Y}}] = \mathbb{E}\left[\left(\tilde{\mathrm{Y}} - c_0\right)^2\right] = \sum_{\substack{\alpha\in\mathcal{A}\\ \alpha\neq 0}} c^2_\alpha. \tag{3.49}$$

Higher-order moments, such as skewness and kurtosis, may also be computed. However, numerical computational methods such as quadrature must be employed to estimate their values since they require the expectation of products of three and four multivariate polynomials, respectively. More details can be found in Sudret [2008].

Global Sensitivity analysis Various sensitivity analysis approaches have been developed in recent years. For an extensive review of different techniques, see Iooss and Lemaître [2015]. Here, I leverage the connection of polynomial representation to global sensitivity measures and use the so-called *Sobol indices* [Sobol', 1993]. These indices are derived from a variance decomposition of model outputs in terms of contributions of each input parameter or combinations thereof. Using Sobol decomposition, one can describe the total variance of model responses in terms of the sum of the summands' variances. This variance decomposition is extensively explained in Sudret [2008].

Sudret [2008] also derives PC-based Sobol indices by leveraging the orthonormality of the polynomial chaos basis. The idea behind these indices is as follows: once the PC representation of the model in Equation (3.15) is available, the expansion coefficients

$c_{\boldsymbol{\alpha}}$ are simply gathered according to the dependency of each basis polynomial, square-summed and normalized, which can be written as

$$S_i^{\mathrm{PCE}} = \frac{\sum\limits_{\boldsymbol{\alpha}\in\mathcal{A}_i} c_{\boldsymbol{\alpha}}^2}{\sum\limits_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}}^2}, \quad \mathcal{A}_i = \{\boldsymbol{\alpha}\in\mathcal{A} : \alpha_i > 0, \alpha_{j\neq i} = 0\} \ . \tag{3.50}$$

In the above equation, S_i^{PCE} is the Sobol index that indicates what fraction of the total variance of the response quantity can be traced back to the joint contributions of the parameters i.

A complementing measure for sensitivity analysis is the *Sobol Total Index*. It expresses the total contribution to the variance of model output due to the uncertainty of an individual parameter in all cross-combinations with other parameters, which reads

$$S_i^{\mathrm{T,PCE}} = \frac{\sum\limits_{\boldsymbol{\alpha}\in\mathcal{A}_i^{\mathrm{T}}} c_{\boldsymbol{\alpha}}^2}{\sum\limits_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}}^2}, \quad \mathcal{A}_i^{\mathrm{T}} = \{\boldsymbol{\alpha}\in\mathcal{A} : \alpha_i > 0\} \ , \tag{3.51}$$

where $S_i^{\mathrm{T,PCE}}$ is simply a summation of all Sobol indices in which the variable i appears as univariate as well as joint influences. The total Sobol index can take values larger than 1 when the impact of interactions among parameters on the total output variance is not negligible. This characteristic is particularly prominent in highly nonlinear problems that are common in engineering applications.

As explained in Section 3.1.5, many computational models provide temporal and/or spatial responses, possibly for more than one SRQ. When PCA is used to reduce the dimensionality of SRQs, the Sobol indices of the principal components need to be computed first. Then, one needs to relate these indices to those for the model outputs. Nagel et al. [2020] establish this link for the models with multivariate outputs. The variance of this conditional expectation reads

$$\mathrm{Var}\left[\mathbb{E}\left[\tilde{\mathrm{Y}}_t \mid \mathrm{X}_i\right]\right] = \sum_{p=0}^{N} \mathrm{Var}\left[\mathbb{E}\left[\tilde{\zeta}_p \mid X_i\right]\right] \bar{\Phi}_{p,t}^2 + 2\sum_{p<q} \mathrm{Cov}\left[\mathbb{E}\left[\tilde{\zeta}_p \mid X_i\right], \mathbb{E}\left[\tilde{\zeta}_q \mid X_i\right]\right] \bar{\Phi}_{p,t}\bar{\Phi}_{q,t}, \tag{3.52}$$

where $\mathrm{Cov}\left[\mathbb{E}\left[\tilde{\zeta}_p \mid X_i\right], \mathbb{E}\left[\tilde{\zeta}_q \mid X_i\right]\right]$ denotes the covariances of the conditional expectations $\mathbb{E}\left[\tilde{\zeta}_p \mid X_i\right]$ and $\mathbb{E}\left[\tilde{\zeta}_q \mid X_i\right]$ for $p, q = 1, \cdots, N$ with $p \neq q$. These expectations

can be calculated by

$$\mathbb{E}\left[\tilde{\zeta}_p \mid X_i\right] = \boldsymbol{c}_{p,\mathbf{0}} + \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\{i\}}} \boldsymbol{c}_{p,\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\mathbf{X}). \tag{3.53}$$

One can now relate the first-order Sobol' index of the output $\tilde{\mathrm{Y}}$ with respect to the input X_i to the corresponding indices of the principal components $\tilde{\zeta}_p$ for $p = 0, \cdots, N$. By dividing Equation (3.52) by the total variance $\mathrm{Var}\left[\tilde{\mathrm{Y}}\right]$, one obtains the first-order Sobol' index with respect to X_i as follows:

$$S_i^t = \frac{\mathrm{Var}\left[\mathbb{E}\left[\tilde{\mathrm{Y}}_t \mid \mathrm{X}_i\right]\right]}{\mathrm{Var}\left[\tilde{\mathrm{Y}}_t\right]} = \sum_{p=0}^{N} S_i^{\tilde{\zeta}_p} \frac{\mathrm{Var}\left[\tilde{\zeta}_p\right]}{\mathrm{Var}\left[\tilde{Y}_t\right]}\phi_{p,t}^2 + 2\sum_{p<q} \frac{\mathrm{Cov}\left[\mathbb{E}\left[\tilde{\zeta}_p \mid X_i\right], \mathbb{E}\left[\tilde{\zeta}_q \mid X_i\right]\right]}{\mathrm{Var}\left[\tilde{\mathrm{Y}}_t\right]}\phi_{p,t}\phi_{q,t} \tag{3.54}$$

where $S_i^{\tilde{\zeta}_p}$ denotes the Sobol index of the principal component $\tilde{\zeta}_p$ with respect to an input variable X_i. This value can be readily determined from the PCE coefficients as follows

$$S_i^{\mathrm{PCE},\tilde{\zeta}_p} = \frac{\sum_{\boldsymbol{\alpha}\in\mathcal{A}_i} c_{\boldsymbol{\alpha}}^2}{\sum_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}}^2}, \quad \mathcal{A}_i = \{\boldsymbol{\alpha} \in \mathcal{A} : \alpha_i > 0, \alpha_{j\neq i} = 0\}\ . \tag{3.55}$$

The terms of covariance in Equation (3.54), given the orthogonality of the polynomial basis, can be written as follows

$$\mathrm{Cov}\left[\mathbb{E}\left[\tilde{\zeta}_p \mid X_i\right], \mathbb{E}\left[\tilde{\zeta}_q \mid X_i\right]\right] = \sum_{\alpha\in\mathcal{A}_{\{i\}}} \boldsymbol{c}_{p,\alpha}\boldsymbol{c}_{q,\alpha}. \tag{3.56}$$

Note that using bootstrap-BaSaPCE allows us to compute credible intervals for the quantities introduced above. Every design from a family of b experimental design results in moments and Sobol indices collections. For example, the empirical confidence interval for the sensitivity analysis with Sobol indices is $\hat{S}_{i[\alpha/2]} \leq S_i \leq \hat{S}_{i[1-\alpha/2]}$ using the α and $1-\alpha/2$ empirical quantiles for significance level $\alpha = 0.05$ [Dubreuil et al., 2014].

3.2 Experimental design

The computational cost of constructing a surrogate model and its accuracy crucially depends on the number of required evaluations of the computationally expensive model in the experimental design (ED) . This is a set of training samples from the joint distribution of the input parameters. Properly designed ED has proved vital for the simultaneous reduction of the effect of noise and bias errors which can raise confidence in the task of Bayesian analysis. This has motivated researchers to examine assorted strategies for constructing the training set $\{\Psi_{\boldsymbol{\alpha}}(\mathrm{X}_i)\}_{i=1}^{N_{\mathrm{ED}}}$ beyond the standard MC sampling. In this context, the influence of different experimental designs on predictions have been adequately addressed in the literature [Simpson et al., 2001, Giunta et al., 2003, Simpson et al., 2004, Queipo et al., 2005, Fajraoui et al., 2017, Hadigol and Doostan, 2018].

In general, the sampling approaches can be categorized into two groups: *classical sampling* and *sequential sampling*. The common practice in classical sampling is to choose the experimental design P grounded only in the information available before any model evaluation, e.g., noise, the relevance of the input parameters, and measurement precision. Then, the computational model is evaluated on the selected samples in the ED, and the surrogate model is finally created. This approach is also known as the one-shot approach, as all the sample points in the ED are specified at once, and no later evaluations of additional samples are made. The *a priori* selection of ED is a fairly challenging task since the determination of optimal sample size is hindered by the lack of prior knowledge about the model behavior.

To tackle this problem, flexible sequential sampling strategies have been proposed, which sequentially determine the samples in the design using the information from previous iterations. The sequential sampling approaches can be grouped into two categories: *sequential space-filling sampling* and *sequential adaptive sampling*. Space-filling approaches make sure that the generated samples cover the entire domain evenly. These sampling approaches are usually developed from some one-shot sampling criteria by selecting the training parameter sets in a sequential manner [Fajraoui et al., 2017]. However, adaptive sequential sampling, also known as adaptive sampling and active learning in machine learning [Settles, 2009], makes more informed choices of samples via the surrogate model itself or data that it learns from, and hence, achieves better

performance with fewer samples than the space-filling sampling, resulting in saving the simulation cost of expensive computational models [Liu et al., 2018].

Here, we adopt a sequential adaptive sampling experimental design (SAED). Algorithm 4 summarizes a typical SAED method. Firstly, an initial batch of samples is selected via a one-shot experimental design. This design can be produced by a common design of experiments approach, such as Latin hypercube sampling [McKay et al., 2000] or random sampling. Next, the model is evaluated provided by the previously selected samples. Then, the surrogate model is trained to construct a relationship between the ED and the quantities of interest. After constructing the surrogate model, its accuracy is estimated using an error metric, e.g., validation error or Leave-one-out error (Section 3.2). Since the initial ED is chosen to be small, the estimated error metric most likely indicates that the ED needs to be enriched.

Algorithm 4: A typical sequential design method

Result: Enriched experimental design

$P \leftarrow$ initial experimental design;
Evaluate the computational model at D;
Train the surrogate model;
Compute the error metric;
while *error* $>$ *prescribed error and No. runs* $<$ *Total No. runs* **do**
    Select new sample P_{new} using sequential design strategy;
    Evaluate the computational model at $\mathbf{d}^\star$;
    $D \leftarrow D \cup \mathbf{d}^\star$;
    Train the surrogate model;
end

A sampling strategy based on active learning selects these additional samples. Some of these strategies will be explained later in detail. Finally, a new surrogate model is built using all the data gathered thus far, and the model accuracy is estimated again. If the surrogate model's prescribed accuracy level or the total number of samples is still not reached, the entire sample selection process is repeated.

Through the sequential selection of samples, more information is available to improve sampling compared to the classical design of experiments [Crombecq, 2011]. The ultimate goal of this algorithm is to reduce the overall number of samples, as evaluating the

samples (running the simulation) is the dominant cost in the entire surrogate modeling process, especially for computationally demanding models. Hence, when the computational bottleneck is the evaluation of the SRQ for any given realization, the additional computational cost of constructing an optimal design via SAED is justifiable.

In what follows, the learning strategies (design criteria) for sequential enrichment of the ED are introduced. The criteria discussed in what follows can be grouped into two categories. The first group can be used for cases where no data is involved, and the goal is to train a surrogate model for uncertainty quantification. The second category of design criteria can be implemented for emulators being trained for model calibration, i.e., the measurement data is available and can assist in the task of Bayesian model validation.

3.2.1 Learning strategies for SAED

To efficiently improve the overall accuracy of a surrogate model, Deschrijver et al. [2011] and Liu et al. [2016a] suggest that a SAED approach must maximize two conflicting parts, namely *local exploitation* and *global exploration*. While local exploitation assists in finding regions where most information can be extracted, global exploration ensures that informative regions that have not been detected yet can be explored. A sampling approach with only exploitation is often biased since it has an imperfect view of the entire domain. To deal with this issue, the global exploration term based on, e.g., some distance criteria is required.

Generally, the new point selected in each sampling iteration given a one-by-one selection process can be done by maximizing the following score function:

$$\mathbf{d}^{\star}_{new} = \underset{d^{+}\in\mathcal{D}}{\operatorname{argmax}}\, Score(global(d^{+}), local(d^{+})), \tag{3.57}$$

where $global(d^{+})$ and $local(d^{+})$ represent the global exploration term and the local exploitation, respectively. In what follows, we will introduce these two terms. These explanations are followed by introducing a trade-off scheme aiming to maintain a balance between local exploitation and global exploration.

Local exploitation

Various adaptive sampling (learning) approaches exist in the literature. These approaches select the new design points based on either an augmented basis or the current trained surrogate model. Fajraoui et al. [2017] provide a brief review and interpretation of a major class of augmented basis-based ODE, known as alphabetic optimal design. Hampton and Doostan [2016] suggest a coherence-optimal sampling in the context of PCE. Liu et al. [2018] present a comprehensive survey of different adaptive sampling approaches according to the sampling criteria for identifying informative regions in the parameter space using the already trained surrogate models. This survey includes all other types of surrogate models, such as the Kalman filter and Gaussian process emulator. In the context of PCE, Ji et al. [2008] and Seeger and Nickisch [2008] propose a criterion that minimizes the differential entropy of the posterior distribution of the coefficients in a Bayesian regression setting. Zhou et al. [2019] also suggest a strategy based on approximations to the expected quadratic loss function, which means the mean squared error.

In what comes next, I investigate two general categories of design criteria: *Active learning* (AL) and *Bayesian active learning* (BAL). Mathematical frameworks of different design criteria (learning strategy) for each category are presented in what follows. The objective of all these criteria is to identify a new design point to run the original computational model, among some prospective designs, which maximizes the expected utility, taking into account the surrogate model at hand and/or the available measurement data for the task of inference. Moreover, these strategies are investigated with an analytical example.

Active learning Following Beck and Guillas [2016], I employ the framework of design and analysis of computer experiments (DACE) proposed by Sacks et al. [1989]. Using DACE, the computational model's output is modeled as a realization of a random field that is typically assumed to be Gaussian. BasaPCE provides Gaussian random fields obtained by Equation (3.31). Using the predicted random fields, one can take advantage of information collected during the experimental design process to determine the next optimal design at which the data need to be collected to refine the surrogate. This procedure is also known as active learning. Active learning criteria are based on

the predictive variance. In what follows, I will introduce two design criteria: active learning MacKay (ALM) and the expected improvement for global fit (EIGF).

Active learning MacKay This strategy was proposed by MacKay [1992b] for active data selection using information-based objective functions and is one of the popular design strategies. At each iteration in the sequential adaptive sampling, ALM selects the design point $\mathbf{d}^\star$ that maximizes the predictive variance of the surrogate:

$$\mathbf{d}^\star_{\mathrm{AL}} = \underset{d^+ \in \mathcal{D}}{\operatorname{argmax}}\, \mathrm{ALM} = \underset{d^+ \in \mathcal{D}}{\operatorname{argmax}}\, \hat{\sigma}^2_D(d^+). \tag{3.58}$$

Here, $\hat{\sigma}^2_D(d)$ is the predictive variance of the surrogate model trained with the current design D evaluated at the candidate design d^+. According to Beck and Guillas [2016], ALM places many design points at the boundary of the design region, i.e., input parameter space. However, boundary points are generally considered less informative than nearby interior points [Krause et al., 2008].

Expected improvement for global fit Lam [2008] proposed EIGF as a variance-based adaptive sampling strategy, which is a modified version of the expected improvement criterion proposed by Jones et al. [1998]. The expected improvement over the nearest observed point can be computed by

$$\mathbf{d}^\star_{\mathrm{AL}} = \underset{d^+ \in \mathcal{D}}{\operatorname{argmax}}\, \mathrm{EIGF} = \underset{d^+ \in \mathcal{D}}{\operatorname{argmax}}\, \left[\mathcal{M}^{\mathrm{PC}}_D(d^+) - \mathcal{M}_D(d')\right]^2 + \hat{\sigma}^2_D(d^+), \tag{3.59}$$

where the first term on the right-hand side tends to a large value when the surrogate response is significantly different from the model response for the nearest observed point d'. The second term, however, accounts for the uncertainty of the surrogate model at the candidate design d^+. Simply put, the EIGF criterion hints toward the region in the parameter space, where either the difference between the prediction and the output at the nearest known design point or the predictive variance is significant. In general, ALM outperforms EIGF, although, for nonstationary output, EIGF could be competitive [Kupresanin and Johannesson, 2011]. Generally, the EIGF is only effective in special cases, especially when the output is constant except for a small region in the design space [Maljovec et al., 2013].

Bayesian active learning BAL was first introduced by Oladyshkin et al. [2020]. The authors present learning strategies (design criteria) that leverage the connection between Bayesian inference and information theory to identify a new training set for the iterative refinement of surrogate models with the Gaussian Process Emulator. Similar to GPE, BaSaPCE provides predictions as a mean value $\mu_{\mathbf{y}}(\theta, x, y, z, t)$ and standard deviation $\sigma_{\mathbf{y}}(\theta, x, y, z, t)$, as discussed in Section 3.1.2. Therefore, initial knowledge on model response $\mathbf{y}(\theta, x, y, z, t)$ in each point of space (x, y, z) and time t for the given exploration parameter set d^+ from the design space $\mathcal{D}$ is encoded in the Gaussian prior probability distribution $\mathcal{N}(\mu_{\mathbf{y}}(d^+, x, y, z, t), \sigma_{\mathbf{y}}(d^+, x, y, z, t))$. Thus, the prior probability distribution of model response $\mathbf{y}(\theta, x, y, z, t)$ for the given candidate parameter set d^+ is forming response space Y that is a multivariate Gaussian, denoted as $\mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})$. According to the Bayesian theorem (Section 2.1.2), we can obtain a posterior probability distribution $p_{d^+}(\mathbf{y}|\mathcal{Y})$ of the model response for the given parameter set d^+, incorporating the observed data $\mathcal{Y}$:

$$p_{d^+}(\mathbf{y}|\mathcal{Y}) = \frac{p_{d^+}(\mathcal{Y}|\mathbf{y})\mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})}{p_{d^+}(\mathcal{Y})}, \tag{3.60}$$

where the term $p_{d^+}(\mathcal{Y}|\mathbf{y})$ is the likelihood function that quantifies how well the surrogate model predictions $\mathbf{y}(d^+, x, y, z, t)$ drawn from the multivariate Gaussian $\mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})$ match the observed data $\mathcal{Y}$. The term $p_{d^+}(\mathcal{Y})$ denotes the Bayesian model evidence value for the given parameter set d^+.

Assuming independent and Gaussian distributed measurement errors, the likelihood function $p_{d^+}(\mathcal{Y}|\mathbf{y})$ can be written as:

$$p_{d^+}(\mathcal{Y}|\mathbf{y}) = (2\pi)^{-\mathrm{N_{out}}/2}|\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\big(\mathcal{Y} - \mathbf{y}(d^+, x, y, z, t)\big)^T \mathbf{\Sigma}^{-1} \big(\mathcal{Y} - \mathbf{y}(d^+, x, y, z, t)\big)\right], \tag{3.61}$$

where $\mathbf{y}(d^+, x, y, z, t)$ comes from $\mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})$ and N_{out} denotes the number of measurement points. In the following sections, we will first show how three choices of utility functions will be introduced. These utilities lead to valid measures of *information gain*, *model evidence* and *information entropy*.

Model evidence-based utility As discussed in Section 2.2.2, BME can be regarded as a metric to rank competing models. Here, we leverage this property of BME to identify

the next suitable training point in the parameter space during the sequential design. In each iteration, we compute the BME value for the prospective design point d^+ using the following expression:

$$\mathrm{BME}_{\mathrm{BAL}} \equiv p_{d^+}(\mathcal{Y}) = \int_Y p_{d^+}(\mathcal{Y}|\mathbf{y})\mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})d\mathbf{y}. \tag{3.62}$$

$\mathrm{BME}_{\mathrm{BAL}}$ in Equation (3.62) can be approximated by:

$$\mathrm{BME}_{\mathrm{BAL}} = \mathbb{E}_{\mathcal{N}_{d^+}(\mu_{\mathbf{y}},\sigma_{\mathbf{y}})}\left[p_{d^+}(\mathcal{Y}|\mathbf{y})\right], \tag{3.63}$$

where the term on the right-hand side denotes the expected value $\mathbb{E}_{\mathcal{N}_{d^+}(\mu_{\mathbf{y}},\sigma_{\mathbf{y}})}$ of the likelihood $p_{d^+}(\mathcal{Y}|\mathbf{y})$ over the prior $\mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})$ provided by the surrogate's prediction. Consequently, the next training point for the surrogate model, i.e., $\mathbf{d}^*_{\mathrm{BAL}} \in \mathcal{D}$, can be identified by maximizing the model evidence $\mathrm{BME}_{\mathrm{BAL}}$

$$\mathbf{d}^*_{\mathrm{BAL}} = \underset{d^+\in\mathcal{D}}{\operatorname{argmax}}\, \mathrm{BME}_{\mathrm{BAL}}. \tag{3.64}$$

Information gain utility A utility function based on mutual information is known as one of the most widely used Bayesian design criteria based on relative entropy. This utility function includes Kullback-Leibler divergence (KLD) [Kullback and Leibler, 1951] and seeks to maximize the expected information gain in moving from the multivariate Gaussian prior $\mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})$ to the posterior $p_{d^+}(\mathbf{y}|\mathcal{Y})$ during the learning procedure.

Formally, the relative entropy $\mathrm{KLD}_{\mathrm{BAL}}\left[p_{d^+}(\mathbf{y}|\mathcal{Y}), \mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})\right]$ can be defined for each candidate sampling point d^+ from the parameter space $\mathcal{D}$ as following:

$$\mathrm{KLD}_{\mathrm{BAL}}\left[p_{d^+}(\mathbf{y}|\mathcal{Y}), \mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})\right] = \int_Y \ln\left[\frac{p_{d^+}(\mathbf{y}|\mathcal{Y})}{\mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})}\right] p_{d^+}(\mathbf{y}|\mathcal{Y})d\mathbf{y}. \tag{3.65}$$

Following Oladyshkin and Nowak [2019], one can avoid multidimensional integration in Equation (3.65) by:

$$\mathrm{KLD}_{\mathrm{BAL}}\left[p_{d^+}(\mathbf{y}|\mathcal{Y}), \mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})\right] = -\ln \mathrm{BME}_{\mathrm{BAL}} + \mathbb{E}_{p_{d^+}(\mathbf{y}|\mathcal{Y})}\left(\ln\left[p_{d^+}(\mathbf{Y}|\mathbf{y})\right]\right). \tag{3.66}$$

Therefore, the optimization problem to select the next training point takes the following

form:

$$\mathbf{d}^*_{\mathrm{BAL}} = \underset{d^+ \in \mathcal{D}}{\operatorname{argmax}} \, \mathrm{KLD}_{\mathrm{BAL}} \left[p_{d^+}(\mathbf{y}|\mathcal{Y}), \mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}}) \right]. \tag{3.67}$$

Note that Equation (3.67) depends not only on $\mathrm{BME}_{\mathrm{BAL}}$ values from Equation (3.66), but also on the cross entropy represented by term $\mathbb{E}_{p_{d^+}(\mathbf{y}|\mathcal{Y})} \left(\ln \left[p_{d^+}(\mathcal{Y}|\mathbf{y}) \right] \right)$. This term reflects how informative the likelihood is (for more details, see Oladyshkin and Nowak [2019]). Moreover, we obtain the last term via a rejection sampling technique (Section 2.1.3) using the evaluations from the already trained surrogate model.

Information entropy-based utility Another utility for selecting the next training point in a BAL has its root in information entropy [Shannon, 1948] and is often used in machine learning. Here, we aim to reduce the expected information loss during the sequential design. The information entropy $\mathrm{H}_{\mathrm{BAL}} \left[p_{d^+}(\mathbf{y}|\mathcal{Y}) \right]$ to assess information loss for each parameter set d^+ as the candidate for next training point can be computed by:

$$\mathrm{H}_{\mathrm{BAL}} \left[p_{d^+}(\mathbf{y}|\mathcal{Y}) \right] = - \int_Y \ln \left[p_{d^+}(\mathbf{y}|\mathcal{Y}) \right] p_{d^+}(\mathbf{y}|\mathcal{Y}) d\mathbf{y}. \tag{3.68}$$

According to Oladyshkin and Nowak [2019], information entropy in Equation (3.68) can be written as following:

$$\begin{aligned} \mathrm{H}_{\mathrm{BAL}} \left[p_{d^+}(\mathbf{y}|\mathcal{Y}) \right] = \ln \mathrm{BME}_{\mathrm{BAL}} &- \mathbb{E}_{p_{d^+}(\mathbf{y}|\mathcal{Y})} \left(\ln \left[\mathcal{N}_{d^+}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}}) \right] \right) \\ &- \mathbb{E}_{p_{d^+}(\mathbf{y}|\mathcal{Y})} \left(\ln \left[p_{d^+}(\mathcal{Y}|\mathbf{y}) \right] \right). \end{aligned} \tag{3.69}$$

We obtain all terms in Equation (3.69) using prior- or posterior-based sampling on the surrogate model's prediction, avoiding any multidimensional integration using methods such as rejecting sampling (Section 2.1.3). Therefore, the optimization problem takes the following form:

$$\mathbf{d}^*_{\mathrm{BAL}} = \underset{d^+ \in \mathcal{D}}{\operatorname{argmin}} \, \mathrm{H}_{\mathrm{BAL}} \left[p_{d^+}(\mathbf{y}|\mathcal{Y}) \right], \tag{3.70}$$

in that, we seek to identify the parameter set $\mathbf{d}^*_{\mathrm{BAL}}$ from the design space $\mathcal{D}$ that corresponds to minimum of information entropy $\mathrm{H}_{\mathrm{BAL}} \left[p_{d^+}(\mathbf{y}|\mathcal{Y}) \right]$.

Global exploration

As alluded to earlier, a pure exploitation criterion for adaptive sampling is often biased. To address this issue, we employ a space-filling-based global exploration criterion in addition to local exploitation to be able to discover the unexplored informative regions in the parameter space. This exploration can be performed using Global Monte Carlo methods. In the context of SAED, one uses these methods to generate many random candidate samples in the parameter (design) space, then computes a score based on a criterion, and finally selects the sample with the best score. In this dissertation, I employ a space-filling input-based method for computing the global exploration scores of design candidates in Equation (3.57). A space-filling design generates an experimental design D in which the samples in the training set are distributed evenly over the design space.

Let us assume a d-dimensional experimental design $D = \{\mathbf{d}_1, \cdots, \mathbf{d}_m\}^T$ containing m samples $\mathbf{d}_i = \left(\mathbf{d}_i^1, ..., \mathbf{d}_i^d\right)$ in the (hyper)cube $[-1, 1]^d$. To achieve an acceptable space-filling design strategy, a candidate sample must maximize certain criteria. I employ a so-called *Intersite-projected distance criterion*, first introduced by Crombecq [2011]. This criterion maximizes two properties, namely the max-min distance and projected distance, by solving a multi-objective optimization problem. The former maximizes the smallest Euclidean distance of any set of points in the design. The latter, however, ensures that if the design points from dimension d are projected to a $(d-1)$-dimensional space along one of the axes, no two points are ever projected onto each other. The objective function that gives scores to a new candidate point d^+ when it is added to an existing design D can be cast as:

$$dist(D, d^+) = \frac{\sqrt[d]{(m+1)} - 1}{2} \underset{\mathbf{d}_i \in D}{\operatorname{argmin}} \sqrt{\sum_{k=1}^{d} |\mathbf{d}_i^k - d^{+,k}|^2} + \frac{m+1}{2} \underset{\mathbf{d}_i \in D}{\operatorname{argmin}} ||\mathbf{d}_i - d^+||_{+\infty} \tag{3.71}$$

Trade-off between local exploitation and global exploration

Recall that we define our sampling objective function in Equation (3.57) as a combination of two competing parts: local exploitation and global exploration, in pursuit of

locating the informative regions and avoiding the formation of clusters in specific regions. To improve sampling performance, we seek a trade-off between local exploitation and global exploration using:

$$S_T = w_l \times S_l + w_g \times S_g \tag{3.72}$$

where S_l and S_g are the local exploitation and the global exploration scores, respectively, as presented earlier. w_l and w_g denote the weights of the criteria and shall satisfy $w_l + w_g = 1$. A valid trade-off not only conducts effective local exploitation in the located region but effectively offers the search algorithm also to consider undetected areas [Liu et al., 2016a].

Various trade-off strategies between local exploitation and global exploration exist in the literature. Liu et al. [2018] classify these strategies into three categories: decreasing strategy, switch strategy, and adaptive strategy. In a decreasing strategy, the global exploration starts with weight w_g close to one to explore the entire domain. With increasing sampling iterations, however, w_g decreases, and the local exploitation receives higher importance weight w_l to identify interesting regions, which means the process converges with a w_l close to one. This strategy has been employed in Kim et al. [2009], Singh et al. [2013] and Turner et al. [2007]. The problem of ignoring global exploration in a decreasing strategy for the final sampling stage led to the development of switch or greedy strategies. Singh et al. [2013] suggested this strategy by switching between exploitation and exploration using a threshold value ε. First, one picks a random value. If this value is smaller than the threshold value ε, the sampling only occurs via pure exploration; otherwise, it changes to pure local exploitation. A generalized version of the switch strategy, namely the adaptive strategy, has been developed to adaptively switch between exploitation and exploration by comparing gained information among successive iterations. Singh et al. [2013] found that the adaptive strategy provides the best results among all strategies.

In this dissertation, I adopt an adaptive trade-off scheme. This approach is based on comparing the model error for the sample selected in the last iteration with its leave-one-out error as described in Section 3.1.3. This method may hint us toward a balance factor w_l. Note that the LOOCV error ε_{LOO} is an estimation of the true error $\varepsilon_{\text{True}}$. The accuracy of this estimation plays a major role in the performance of the sampling process. It is commonly the case that LOOCV error over-/under-estimate the actual

error [Viana et al., 2009, Liu et al., 2016b]. Consequently, the local exploitation score becomes biased. To avoid this, we will assign a higher weight to the global exploration for the iterations in which the difference between the LOOCV error ε_{LOO} and the true error $\varepsilon_{\text{True}}$ is significant. To this end, the factor w_g in Equation (3.72) can be derived from the idea of adjusting ε_{LOO} so that it is a good representation of the true error $\varepsilon_{\text{True}}$.

Let i be the iteration number, we have selected a new training sample $\mathbf{d}_{i-1}$ in the $(i-1)$-th iteration. Moreover, assume that $\mathbf{y}_{i-1}$ denotes the corresponding true model response using the simulator. The acquired information in the $(i-1)$-th iteration can help determine the w_g value for the i-th iteration. A dynamic adjustment of w_g can be performed by comparing the estimated LOOCV error $\varepsilon_{\text{LOO}}(\mathbf{d}_{i-1})$ and the true prediction error $\varepsilon_{\text{True}}(\mathbf{d}_{i-1})$. Thus, the global weight w_g takes the following form

$$w_g = \begin{cases} 0.5, & q = 1 \\ \min\left[0.5 \times \frac{\varepsilon^2_{\text{True}}(\mathbf{d}_{i-1})}{\varepsilon^2_{\text{LOO}}(\mathbf{d}_{i-1})}, 1\right], & q > 1 \end{cases} . \tag{3.73}$$

The balance factor w_g receives 0.5 in the first iteration, meaning that the local exploitation and the global exploration are treated equally. For iteration $i > 1$, once $\mathbf{d}_{i-1}$ is obtained, one builds the surrogate $\hat{\mathcal{M}}_{i-1}$. Then, the LOOCV error $\varepsilon_{\text{LOO}}(\mathbf{d}_{i-1})$ is obtained using Equation (3.38) and the true prediction error $\varepsilon_{\text{True}}(\mathbf{d}_{i-1})$ is a root mean square error. Figure 3.6 illustrates how w_g changes with respects to $\varepsilon_{\text{True}}/\varepsilon_{\text{LOO}}$.

The following observation can be made from Figure 3.6:

- if $\varepsilon^2_{\text{True}} > \varepsilon^2_{\text{LOO}}$, the LOOCV error is an underestimation of the true error. Thus, the weight w_g decreases from 0.5 to $0.5 \times \frac{\varepsilon^2_{\text{True}}(\mathbf{d}_{i-1})}{\varepsilon^2_{\text{LOO}}(\mathbf{d}_{i-1})}$, i.e. the global exploration is favored.

- if $\varepsilon^2_{\text{True}} < \varepsilon^2_{\text{LOO}}$, the LOOCV error overestimates the true error. Thus, the local exploitation is favored with a weight $\min\left[0.5 \times \frac{\varepsilon^2_{\text{True}}(\mathbf{d}_{i-1})}{\varepsilon^2_{\text{LOO}}(\mathbf{d}_{i-1})}, 1\right]$.

- $\varepsilon^2_{\text{True}} = \varepsilon^2_{\text{LOO}}$ indicates that the LOOCV error is an accurate estimation of the true error and as a result the local exploitation and the global exploration are treated equally.

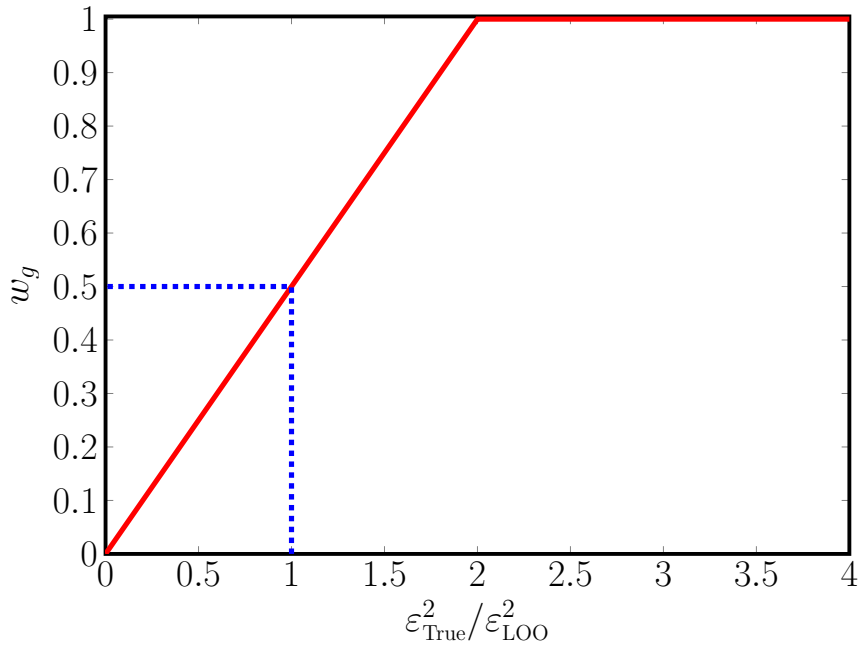


Figure 3.6: The global exploration weight for $i > 1$.

3.2.2 Numerical experiment

SAED for an analytical test case

For the first numerical study, I consider a non-linear analytical function $\boldsymbol{y}(\boldsymbol{\theta}, t)$ with ten ($n = 10$) uncertain parameters $\boldsymbol{\theta} = \{\theta_1, ..., \theta_n\}$, used in Oladyshkin and Nowak [2019] as:

$$\boldsymbol{y}(\boldsymbol{\theta}, t) = \left(\theta_1^2 + \theta_2 - 1\right)^2 + \theta_1^2 + 0.1 \times \theta_1 \exp\left(\theta_2\right) - 2 \times \theta_1 \sqrt{0.5t} + 1 + \sum_{i=2}^{n} \frac{\theta_i^3}{i}, \quad (3.74)$$

where the prior parameter distribution is considered to be independent and uniform with $\theta_i \sim \mathcal{U}(-5, 5)$ for $i = 1, ..., n$. Moreover, I construct a test scenario by generating ten synthetic observed data values $\boldsymbol{y}_\star = \boldsymbol{y}(\boldsymbol{\theta}, t_k)$ with $k = 1, ..., 10$ corresponding to $\theta_i = 0 \ \forall i$.

To assess the prediction accuracy of $\boldsymbol{y}(\boldsymbol{\theta}, t)$, in Equation (3.74) comparing to the synthetic observed data $\boldsymbol{y}_\star$, I use the likelihood function in (2.6), assuming independent and Gaussian distributed error of $\sigma_\varepsilon = 2$. In what follows, I investigate the performance of the SAED with different strategies, with a particular focus on different goals

of employing surrogate models in a Bayesian framework. These goals are posterior reconstruction and assessment of likelihood arguments, such as BME and KLD.

Posterior reconstruction In many cases, the objective of using a surrogate model is to accelerate the Bayesian inference (Section 2.1.3), which yields posterior parameter distributions. This posterior is an update of the prior distribution on the input model parameters after comparing it to the observed data. Here, I analyze a 2D case of the analytical function, i.e., with two parameters θ_1 and θ_2, for the sake of better visualization of the posterior parameter space. As a reference, one can generate a true posterior distribution using the test scenario explained earlier. Next, I will present the surrogate model's posterior distribution during the sequential refinement based on information-based utilities introduced in AL and BAL frameworks (Section 3.2.1).

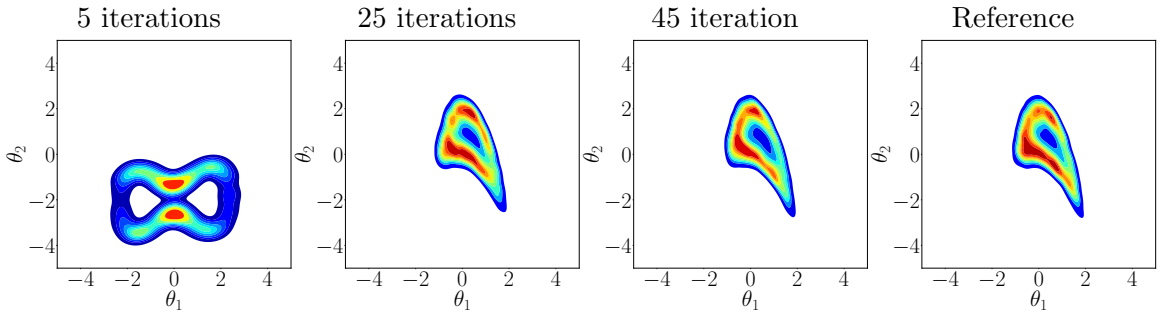


Figure 3.7: The surrogate-based posterior distributions after 5, 25, and 45 sequential learning steps with AL-ALM versus the original posterior for the 2D case of the analytical function using MCMC method (Section 2.1.3)

Figure 3.7 illustrates the surrogate-based posterior distribution after 5, 25 and 45 iterations against the reference posterior distribution. The initial surrogate was trained with two training samples using the LHS method [McKay et al., 2000]. Then, SAED was started with ALM within the active learning framework. Figure 3.8 shows the surrogate-based posteriors using BAL with KLD utility. These figures indicate that both learning strategies retrieve the non-Gaussian posterior parameter distribution reasonably well after a few iterations for the analyzed 2D analytical test case. Additionally, it can be observed that there is no significant difference concerning the posterior distribution after 25 iterations. This observation indicates that the SAED algorithm can effectively refine the surrogate model only with some iterations for this 2D problem.

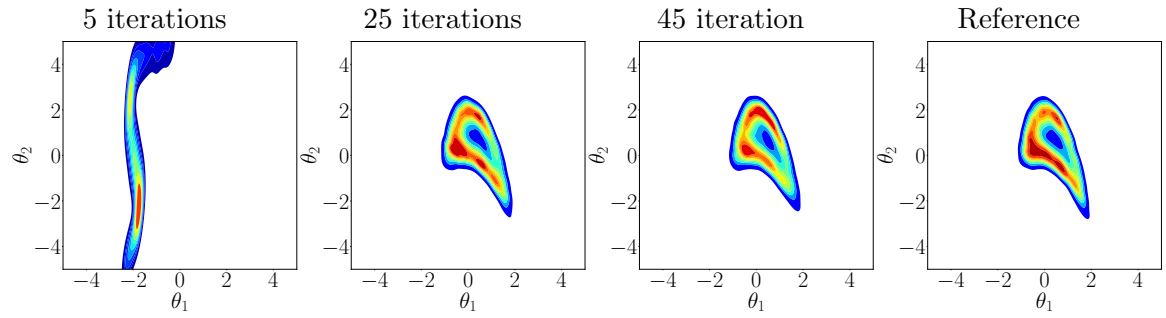


Figure 3.8: The surrogate-based posterior distributions after 5, 25 and 45 sequential learning steps with BAL-KLD versus the original posterior for the 2D case of the analytical function using MCMC method (Section 2.1.3)

Further assessment of the convergence So far, we visually inspect the posterior achieved by surrogates with their respective references generated from the original model for the 2D version of the analytical test case. In what follows, I will explore the convergence of some scores, which include information regarding the likelihood and the posterior during the sequential design for the original 10D problem. As discussed earlier in Section 2.3, a multi-model comparison relies on approximating each model's BME value. A BME value can be approximated by taking the average of the likelihood for the entire prior parameter space in Equation (2.9).

I monitor the changes in surrogate-based BME and Kullback-Leibler divergence between prior and posterior with their reference values during the learning process. Oladyshkin and Nowak [2019] define and approximate these measures as below:

$$\text{BME} = p(\mathbf{d}) = \int_{\Theta} p(\mathbf{d} \mid \theta) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_{p(\boldsymbol{\theta})}(p(\mathbf{d} \mid \boldsymbol{\theta})) \approx \frac{1}{N} \sum_{i=1}^{N} \left(p\left(\mathbf{d} \mid \boldsymbol{\theta}_i\right)\right) \tag{3.75}$$

$$\text{DKL}[p(\boldsymbol{\theta} \mid \mathbf{d}), p(\boldsymbol{\theta})] = -\ln \text{BME} + \frac{1}{N_p} \sum_{i=1}^{N_p} \left(\ln \left[p\left(\mathbf{d} \mid \boldsymbol{\theta}_i\right)\right]\right) \tag{3.76}$$

Figures 3.9 and 3.10 present the metrics in Equations (3.75) and (3.76) compared to their reference values with the increasing number of training samples in SAED. In each iteration of surrogate model refinement, one sample has been added to the ED. For a better visibility, the plots show the results for every 15 iterations. Due to the random nature of the discussed sampling techniques, the results presented in the following are

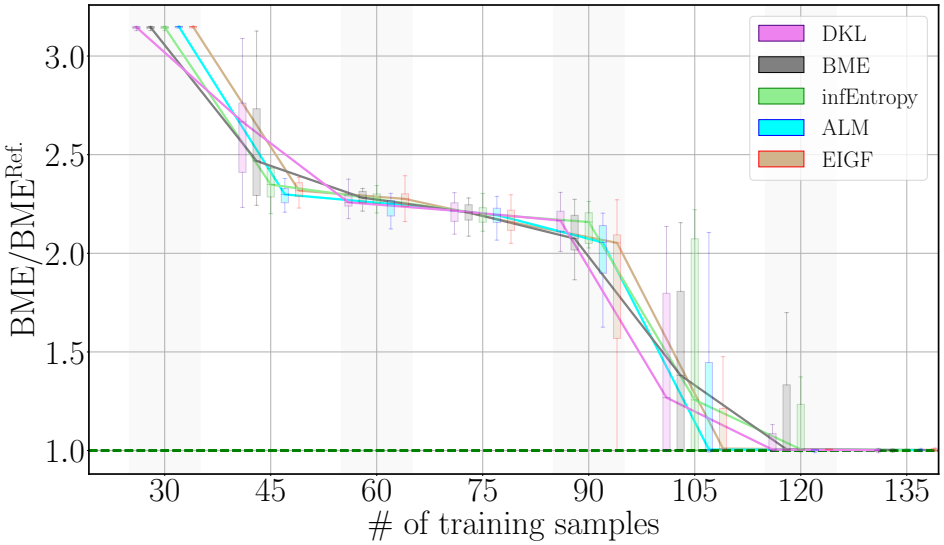


Figure 3.9: Evolution of BME with increasing training samples using BaSaPCE-SAED for the analytical function ($d = 10$, $p = 12$, $q = 0.5$)

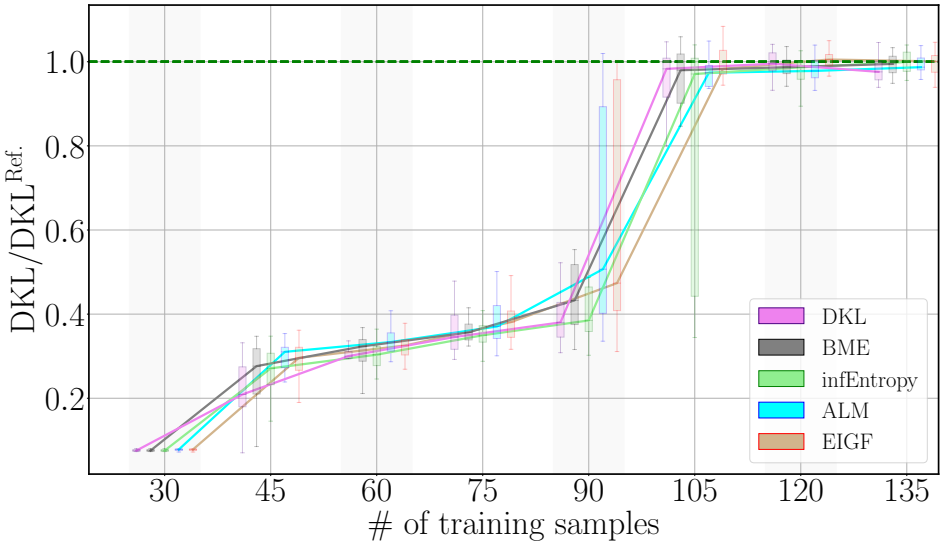


Figure 3.10: Evolution of KLD with increasing training samples using BaSaPCE-SAED for the analytical function ($d = 10$, $p = 12$, $q = 0.5$)

obtained by running 20 independent replications. The variability of the values due to these replications are presented with the box plots. The solid lines show the median of the values resulting from replications.

Figures 3.9 and 3.10 reveal that the BME and DKL values converge to the reference value using BaSaPCE with only 135 model runs for a highly non-linear problem con-

taining the parameter space of 10 dimensions. It can also be noticed that BME values of all design criteria reach convergence after 120 training samples. In terms of BME, the AL design criteria ALM and EIGF reach the convergence even after 75 iterations (105 simulation runs). The KLD evolution with the increasing training samples Figure 3.10 indicates more or less simultaneous convergence.

In this chapter, I addressed the following research question: how can a better surrogate prediction accuracy be achieved for computationally demanding models under time constraint? I investigated two approaches: a) clever selection of the PCE terms using Sparse Learning approaches, and b) use of a sequential adaptive experimental design that makes more informed choices of training samples.

A numerical experiment to compare different sparse learning methods introduced in Section 3.1.2 revealed that the Bayesian Sparse Learning methods showed promising prediction accuracy with increasing model complexity, *i.e.*, number of input parameters. A sequential adaptive learning method was presented to refine the surrogate by selecting the training samples iteratively. This approach strikes a trade-off between exploitation and exploration utility functions. With the help of a numerical example, it was shown how surrogate modeling can benefit from this approach to provide good accuracy to perform surrogate-based model calibration and ranking in a multi-model setting.

The upcoming chapter introduces a Python implementation of the Bayesian multi-model validation Framework described so far. This python package provides an automated surrogate-based sensitivity analysis, Bayesian calibration-validation and multi-model comparison.

4 BayesValidRox: a Python Package for Bayesian Multi-Model Comparison

A Python implementation of the Bayesian multi-model validation Framework described so far is presented in this chapter. The Python package called *BayesValidRox* has been developed as an open-source, object-oriented implementation of this framework. With a modular structure, it provides automated surrogate-based sensitivity analysis, Bayesian calibration-validation and multi-model comparison. The package is available on The Python Package Index (PyPI) software repository*.

Figure 4.1 shows all available modules of *BayesValidRox* and their dependencies. The upcoming sections serve as a tutorial to guide users in using different modules of *BayesValidRox* to perform UQ analysis of their problems at hand. In what follows, I show how to couple computational models to *BayesValidRox*, train surrogate models, and post-process the trained model. Section 4.4 shows how one can perform calibration via a surrogate-based Bayesian inference. Validation and model comparison will be discussed in Section 4.5.

4.1 Model coupling with PyLink

The first step to use *BayesValidRox* is to connect the computational model(s) to the package. Analyses can proceed smoothly once this step has been completed. Computational models can be coupled to *BayesValidRox* in direct and indirect via a python

*https://pypi.org/project/bayesvalidrox/

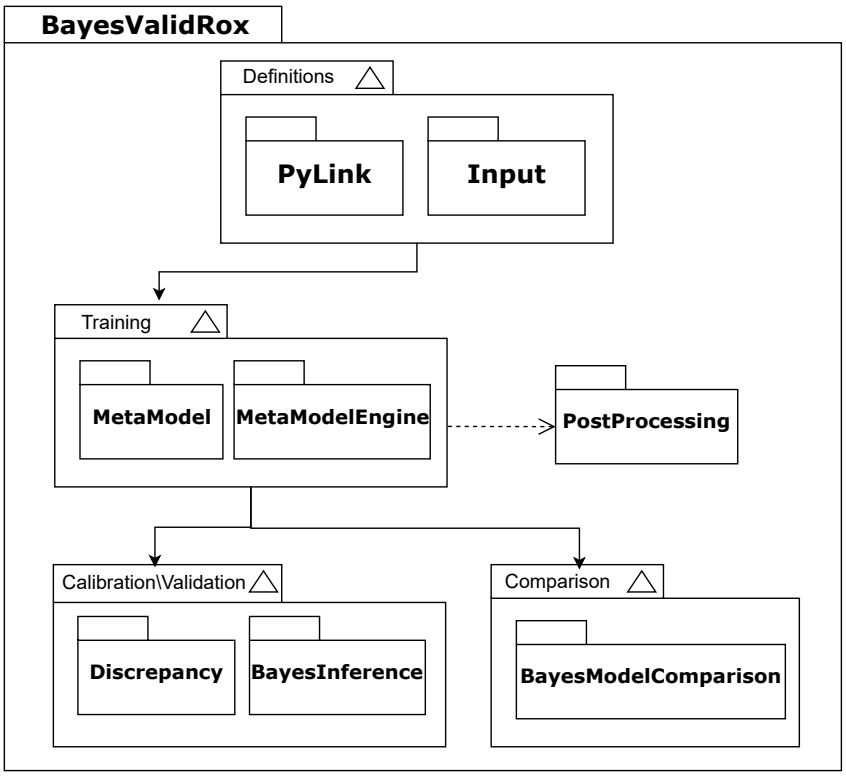


Figure 4.1: The available modules in *BayesValidRox* and their dependencies.

wrapper. To use the option with a python wrapper, after instantiating the object `PyLinkForwardModel`, the attribute `link_type` needs to be set to `'Function'`. Then, we can pass the name of the python wrapper (python file) to the object.

```
from bayesvalidrox import PyLinkForwardModel

Model = PyLinkForwardModel()

Model.link_type = 'Function'  # Link type
Model.py_file = 'AnalyticalFunction'  # Name of the python wrapper
Model.name = 'AnalyticFunc'  # Name of the model
Model.Output.names = ['Z'] # List with the names of the  model ouputs
```

Listing 4.1: How to use `PyLink` when using a Python wrapper

But, what is actually a python wrapper? A wrapper or a binder allows the execution of a third-party software/solver within the scope of *BayesValidRox*. This is basically a function, written in a python file, that takes the parameters in an array of shape `(n_samples, n_params)` as an argument and returns a dictionary with the `x_values` and

output arrays for given output names passed to `Model.Output.names`. Here, `x_values` is an array or a list of the time steps or point IDs corresponding to the outputs under investigation. For multi-output cases, each output is expected to be saved in the output dictionary with the corresponding name as the dictionary key. The `x_values` are expected in a dictionary. An example of a python wrapper is shown in Listing 4.2.

```
def python_wrapper(xx):

    # Extract No. of samples and parameters
    n_samples, n_params = xx.shape

    # Prepare the time steps
    t = np.arange(0, 10, 1.) / 9

    # Compute output
    term1 = (xx[:, 0]**2 + xx[:, 1] - 1)**2
    term2 = xx[:, 0]**2
    term3 = 0.1 * xx[:, 0] * np.exp(xx[:, 1])
    outputs = term1 + term2 + term3 + 1

    # Prepare output dictionary
    output_dict = {
        'x_values': t,
        'Z': outputs
    }
    return output_dict
```

Listing 4.2: An example of a Python wrapper

As for the second option, we need to pass `'PyLink'` to the class `obj.type_name` and give the command to run the model, path to the input template files, and a python file to parse the output file written by the third-party software in a file. Let us assume a model named `Beam9points` that can be run by an executable with the same name to be found in the current directory, hence `exe_path = os.getcwd ()` (this the default value - change if the execution file is not in the same directory) and a text-based input file `SSBeam_Deflection.inp`. To run this model, we need to execute the command `myBeam9points SSBeam_Deflection.inp`. This action starts a simulation and the outputs `Deflection [m]` are saved in a text based `SSBeam_Deflection.out` file.

```
Model = PyLinkForwardModel()

```

```
Model.link_type = 'PyLink'
Model.name = 'Beam9points'
Model.input_file = "SSBeam_Deflection.inp"
Model.input_template = "SSBeam_Deflection.tpl.inp"
Model.shell_command = "myBeam9points SSBeam_Deflection.inp"
Model.exe_path = os.getcwd()
Model.Output.parser = 'read_Beam_Deflection'
Model.Output.names = ['Deflection [m]']
Model.Output.file_names = ["SSBeam_Deflection.out"]
```

Listing 4.3: PyLink to use a third-party solver

After execution of each simulation, the output values are read by means of a parser python script, whose name is passed to `Model.Output.parser`. An example of such a script is presented in Listing 4.4.

```
import numpy as np
def read_Beam_Deflection(file_names):
    # Read the outputs from the output file
    outputs = np.loadtxt(file_names[0], delimiter=',')
    # Prepare output dictionary
    output_dict = {
        'x_values': np.arange(0, 5.6, 5./9),
        'Deflection [m]': outputs
    }
    return output_dict
```

Listing 4.4: An example of a parser script to read the simulation outputs

In addition to `obj.input_file`, we need to create a template for the input file with the naming convention `[input_file_name].tpl.[input_file_extension]`. *BayesValidRox* copies the template files to the dedicated directories and renames them according to the simulation run number. It also searches for the keywords with `<Xi>` pattern and replaces the parameter set values. Here, `i` represents the parameter number defined later in the input section. An example of a template input file is presented in Listing 4.5.

```
% Input file for the simply supported beam model
<X1> % b in m
<X2> % h in m
5 % L in m
<X3> % E in Pa
```

```
<X4> % p in N/m
```

Listing 4.5: An example of a template input file

Observation data To perform uncertainty-aware calibration and/or validation, the observation (measurement) data must be defined. This can be done in two ways: direct and indirect, as shown in Listing 4.6. The former can be realized by passing the observation data in the form of a dictionary to `obj.observations` and `obj.observations_valid` for calibration and validation, respectively. The indirect option is to read the data from a text-based file, given to `obj.meas_file` or `obj.meas_file_valid` for the validation step. These text-based files must include data for each output with the column headers given by `Model.Output.names`.

```
# Direct way
Model.observations = {}
Model.observations['Time [s]'] = np.arange(0, 10, 1.) / 9
Model.observations['Z'] = np.repeat([2.], 10)

# Indirect way
Model.meas_file = 'MeasuredData.csv'
Model.meas_file_valid = 'MeasuredData_Valid.csv'
```

Listing 4.6: An example of how to pass observation data

4.2 Uncertain Input Parameters

So far, we defined all the specifications required to link *BayesValidRox* with a third-party software/solver. The next step is to define the model input parameters. These parameters are selected by the user among all the modeling parameters, whose values are not known *a priori*. The uncertainties of these parameters are defined by their distributions or data gathered from experimental campaigns. In the lack of information on the parameter distribution, uniform distribution with physically meaningful ranges is advisable.

To define the uncertain parameters, one needs to start with instantiating the `Input` class. In the next step, for each parameter, an object needs to be created by calling

`Input.add_marginals()`. This will instantiate an object called `Marginals`. This object accepts a string as `name`, a distribution type as `dist_type`, and distribution parameters as `parameters` variables. In case that the input parameters do not follow specific distribution, one can pass available data directly to the `Input.Marginals` object and to the `input_data` variable.

```
from bayesvalidrox import Input
Inputs = Input()

# ----- First option -----
# Parameter 1
Inputs.add_marginals()
Inputs.Marginals[0].name = '$X_{1}$'
Inputs.Marginals[0].dist_type = 'uniform'
Inputs.Marginals[0].parameters =  [-5, 5]

# Parameter 2
Inputs.add_marginals()
Inputs.Marginals[1].name = '$X_{2}$'
Inputs.Marginals[1].dist_type = 'uniform'
Inputs.Marginals[1].parameters =  [-5, 5]

# ----- Second option -----
# Read data from a file
input_params = np.load('InputParameters.npy')

# Parameter 1
Inputs.add_marginals()
Inputs.Marginals[0].name = '$X_{1}$'
Inputs.Marginals[0].input_data = input_params[:, 0]

# Parameter 2
Inputs.add_marginals()
Inputs.Marginals[1].Name = '$X_{2}$'
Inputs.Marginals[1].input_data = input_params[:, 1]
```

Listing 4.7: An example of how to define uncertain model parameters

The distribution types implemented in *BayesValidRox* are: `uniform`, `normal`, `lognormal`, `exponential`, `gamma`, `beta`, and `weibul`. The distribution parameters, such as lower bound, upper bound for the `uniform` distribution, mean and standard deviation for the

`normal` distribution, location and scale for `lognormal` distribution need to be passed to `Inputs.Marginals[i].parameters`. If the second option in Listing 4.6 is selected, the distribution is represented by a kernel-density estimate using Gaussian kernels. This estimate is a way to represent the probability distribution function (PDF) of random variables in a non-parameteric manner [Scott, 2015].

4.3 Surrogate Modelling

Up until now, we have only defined the model specifications and the distributions for the input parameters. In this section, I will show how a user can set up and train a surrogate (meta) model with *BayesValidRox*. Moreover, the post-processing step after training will be introduced which can provide useful information about the quality of the surrogate model, sensitivity analysis, and the statistical moments of the simulations.

4.3.1 Training a meta-model

First, the user must instantiate a `MetaModel` object and pass the already specified `Input` and `Model` objects as the arguments. Second, the type of the metamodel must be specified, which can be selected from (classical) polynomial chaos expansion `PCE`, and arbitrary polynomial chaos expansion (`aPCE`). For more details, see Chapter 3.

As discussed in Section 3.1.5, one can employ an output dimensionality reduction technique to treat the spatial and temporal dependencies. This can be specified by `MetaModelOpts.dim_red_method = 'PCA'`. The number of principal components can be defined either directly using `n_pca_components` or a variance threshold `var_pca_threshold`, for which the components can explain the underlying variance of the simulation outputs.

For the `PCE` or `aPCE`, *BayesValidRox* expects some additional parameters for the regression method (`pce_reg_method`), the polynomial degree (`pce_deg`), and q-quasi-norm (`pce_q_norm`) for the hybrid truncation of $\boldsymbol{\alpha}$ in Equation (3.2). The list of available regression methods can be found as a comment in Listing 4.8. Bayesian Sparse aPCE, presented in Section 3.1.2 can be specified using `pce_reg_method = 'FastARD'` or `'BCS'`. It is advised to use the bootstrapping option to account for the variability of the surrogate predictions due to the finite size of the experimental design, see Section 3.1.6.

There are two options available: a) normal and b) fast. While the former uses the selected `pce_reg_method` for all the bootstrapping phase, the latter only performs the regression with the selected regression method at the beginning and uses the information obtained to update the coefficients with an ordinary least-square (OLS) method. The number of the bootstrap sampling iteration can be prescribed using the argument `n_bootstrap_itrs`. The attribute `pce_deg` accepts the polynomial degree either as an integer or as an array containing multiple degrees. For the latter, the algorithm compares the metamodels for all given degrees and selects the one with the lowest LOOCV error as defined by Equation (3.38). The parameter `pce_q_norm` must be between 0 and 1 and has a default value of 1.

```
from bayesvalidrox import MetaModel
MetaModelOpts = MetaModel(Inputs, Model)

# Select if you want to preserve the spatial/temporal dependencies
MetaModelOpts.dim_red_method = 'PCA'
MetaModelOpts.var_pca_threshold = 99.999
# or
MetaModelOpts.n_pca_components = 10

# Select your metamodel method
# 1) PCE (Polynomial Chaos Expansion) 2) aPCE (arbitrary PCE)
MetaModelOpts.meta_model_type = 'aPCE'

# -----------------------------------------------
# ------------- PCE Specification ---------------
# -----------------------------------------------
# Select the sparse least-square minimization method for
# the PCE coefficients calculation:
# 1)OLS: Ordinary Least Square  2)BRR: Bayesian Ridge Regression
# 3)LARS: Least angle regression  4)ARD: Bayesian ARD Regression
# 5)FastARD: Fast Bayesian ARD Regression
# 6)BCS: Bayesian Compressive Sensing, a.k.a. FastLaplace
# 7)OMP: Orthogonal Matching Pursuit
# 8)VBL: Variational Bayesian Learning
# 9)EBL: Emperical Bayesian Learning
MetaModelOpts.pce_reg_method = 'FastARD'

# Bootstraping option
# 1) normal 2) fast
```

```
MetaModelOpts.bootstrap_method = 'fast'
MetaModelOpts.n_bootstrap_itrs = 100

# PCE degree
# pce_deg accepts degree as a scalar or a range.
MetaModelOpts.pce_deg = np.arange(1, 12)

# q-quasi-norm 0<q<1 (default=1)
MetaModelOpts.pce_q_norm = 0.75
```

Listing 4.8: An example of the metamodel definition

Experimental design To produce the training set for creating the metamodel, one needs to add an `ExpDesign` object to the `MetaModel` object. This can be done by `MetaModelOpts.add_ExpDesign()`. *BayesValidRox* offers two methods for the computer experimental design Section 3.2: one-shot (`normal`) or sequential (`sequential`). For both design methods, the number of the sampling points (`n_init_samples`) and sampling method (`sampling_method`) shall be passed to the `ExpDesign` object.

The following sampling methods are available: `random`, `latin_hypercube` [McKay et al., 2000], `sobol` [Sobol', 1967], `halton` [Halton, 1964], `hammersley` [Hammersley, 2013], `chebyshev` [Stewart, 1996], `grid`. The last three options provide full-tensor designs with N^d samples, where N is the number of requested samples and d number of parameters. One can also pass the already existing training set by choosing `user` sampling methd in two formats: a) directly to `MetaModelOpts.ExpDesign.X` and `MetaModelOpts.ExpDesign.Y` or b) via an hdf5 file. When passing the experimental design using the first option, special care must be given to the formats. The training parameter sets must be of size `(n_samples,n_params)`, whereas `MetaModelOpts.ExpDesign.Y` receives a dictionary containing the `x_values` and the model outputs. Using the second option, the hdf5 file shall contain the samples in `EDX/init_` and the corresponding outputs as `EDY/output_name/init_`. The `x_values` for single output cases must be stored under the `x_values` path and for multi-output, they must follow `x_values/output_name` path. Note that after the successful execution of the analysis with *BayesValidRox* the corresponding experimental design is stored in an hdf5 file with the described structure. This approach makes it easy to use the existing experimental design for further analysis.

```
MetaModelOpts.add_ExpDesign()
```

```
2
3 # One-shot (normal) or Sequential Adaptive (sequential) Design
4 MetaModelOpts.ExpDesign.method = 'normal'
5 MetaModelOpts.ExpDesign.n_init_samples = 100
6
7 # Sampling methods
8 # 1) random 2) latin_hypercube 3) sobol 4) halton 5) hammersley
9 # 6) chebyshev(FT) 7) grid(FT)
10 MetaModelOpts.ExpDesign.sampling_method = 'random'
11
12 # Alternative: If the experimental design already exists as a hdf5
13 # file, pass the experimental design object
14 MetaModelOpts.ExpDesign.sampling_method = 'user'
15 # Option I:
16 MetaModelOpts.ExpDesign.X = samples
17 MetaModelOpts.ExpDesign.Y = output_dict
18 # Option II:
19 MetaModelOpts.ExpDesign.hdf5_file = 'ExpDesign_Model_Name.hdf5'
```

Listing 4.9: An example of the experimental design

Sequential experimental design For computationally demanding models, a sequential experimental design can assist in reducing the computational cost of the training by refining the surrogate model using an informed selection of training parameter sets. More details can be found in Section 3.2. *BayesValidRox* offers multiple SAED strategies that can be categorized into three groups: space-filling, variance-based design, and Bayesian active design. The last group requires data and the discrepancy error to operate.

```
1 # ------------------------------------------------
2 # ----- Initial Exp. Design configuration --------
3 # ------------------------------------------------
4 MetaModelOpts.add_ExpDesign()
5
6 # One-shot (normal) or Sequential Adaptive (sequential) Design
7 MetaModelOpts.ExpDesign.method = 'sequential'
8 MetaModelOpts.ExpDesign.n_init_samples = 10
9
10 # Sampling methods
11 MetaModelOpts.ExpDesign.sampling_method = 'latin_hypercube'
```

```
12
13 # ------------------------------------------------
14 # ------- Sequential Design configuration --------
15 # ------------------------------------------------
16 # Set the sampling parameters
17 MetaModelOpts.ExpDesign.n_new_samples = 1 # No. samples in each
       iteration
18 MetaModelOpts.ExpDesign.n_max_samples = 150  # No. of maximum samples
19
20 # Trade-off schemes
21 # 1) None 2) 'equal' 3)'epsilon-decreasing' 4) 'adaptive'
22 MetaModelOpts.ExpDesign.tradeoff_scheme = 'adaptive'
23 # -------- Exploration ------
24 # 1)'Voronoi' 2)'random' 3)'latin_hypercube' 4)'LOOCV' 5)'dual
       annealing'
25 MetaModelOpts.ExpDesign.explore_method = 'random'
26
27 # Use when 'Voronoi' or 'random' or 'latin_hypercube' chosen
28 MetaModelOpts.ExpDesign.n_canddidate = 1000
29 MetaModelOpts.ExpDesign.n_cand_groups = 4
30
31 # -------- Exploitation ------
32 # 1) Space-filling 2) BayesActDesign 3) VarOptDesign
33 MetaModelOpts.ExpDesign.exploit_method = 'BayesActDesign'
34
35 # VarBasedOptDesign -> Active learning
36 # 1)ALM 2)EIGF 3) MI 4) ALC
37 MetaModelOpts.ExpDesign.util_func = 'ALM'
38
39 # BayesActDesign -> Bayesian active learning
40 # 1)DKL 2)BME 3) infEntropy
41 MetaModelOpts.ExpDesign.util_func = 'DKL'
```

Listing 4.10: An example of the sequential experimental design

We can monitor the accuracy of the metamodel via the LOOCV error with increasing training samples. Having some test sets available, the evolution of the validation error can also be examined. In the case of Bayesian calibration, the changes to the BME and KLD are also computed and displayed in the console. However, the measurement data and the discrepancy model (see Section 4.4) must be defined, as shown in Listing 4.11.

```
1 # Defining the measurement error, if it's known a priori
```

```
known_sigma_squared = {
    'Z': np.array([1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0])
    }
DiscrepancyOpts = Discrepancy('')
DiscrepancyOpts.type = 'Gaussian'
DiscrepancyOpts.parameters = known_sigma_squared
MetaModelOpts.Discrepancy = DiscrepancyOpts

# For calculation of validation error in each iteration
MetaModelOpts.valid_samples = valid_set
MetaModelOpts.valid_model_runs = {'Z': valid_outputs}
```

Listing 4.11: An example of the necessary definitions for monitoring convergence metrics in sequential experimental design

Users can plot these convergence measures using a helper function in the postprocessing steps.

After prescribing all the necessary parameters, the training task can be initiated using `create_metamodel` method of the `MetaModel` object. After training, it is advisable to store the metamodel using python's `joblib` module. The `MetaModel` object can be loaded for further analysis later. The code snippet in Listing 4.12 provides an example of the model training and storing/loading the metamodel object.

```
# Train the meta model
meta_model = MetaModelOpts.create_metamodel()

# Train the meta-model for sequential design
from bayesvalidrox import MetaModelEngine
meta_model_engine = MetaModelEngine(MetaModelOpts)
meta_model_engine.run()
meta_model = meta_model_engine.MetaModel

# Save Meta models
with open(f'MetaModel_{Model.name}.pkl', 'wb') as output:
    joblib.dump(meta_model, output, 2)

# Load MetaModel
with open(f'MetaModel_{Model.name}.pkl', 'rb') as input:
    meta_model = joblib.load(input)
```

Listing 4.12: An example of the about the model training and object storage

4.3.2 Post-processing

Once the metamodel is successfully trained, one can use some helper functions provided by the `PostProcessing` module to perform post-processing and accuracy checks.

Moment analysis As discussed in Section 3.1.7, one can compute and visualize the moments from the coefficients using Equations (3.48) and (3.49).

```
from bayesvalidrox import PostProcessing

# Instantiate the post-processing object
PostPCE = PostProcessing(MetaModel, name='calib')

# Compute the moments
pce_means, pce_stds = PostPCE.plot_moments()
```

Listing 4.13: An example of the moment analysis for a PCE-based metamodel

Running the code snippet in Listing 4.13 creates a folder named `Outputs_PostProcessing_calib` and stores the moment plots.

Sensitivity analysis *BayesValidRox* also allows you to perform a sensitivity analysis of the input parameters via Sobol indices based on the theory discussed in Section 3.1.7. Similar to the moment analysis, the calculations are based on the coefficients of the metamodel with no extra model evaluations. Listing 4.14 shows an example of a sensitivity analysis with the `PostProcessing` module. The method `sobol_indices` returns the total indices in a dictionary. Moreover, it generates plots of the Sobol indices in the form of bar plots and line plots for cases with temporal/spatial resolved outputs. For bar plot, set the argument `plot_type` to `'bar'`. The indices in the plots also come with confidence intervals resulting from the bootstrapping.

```
from bayesvalidrox import PostProcessing

# Instantiate the post-processing object
PostPCE = PostProcessing(MetaModel, name='calib')

# Compute the Sobol indices
```

```
total_sobol = PostPCE.sobol_indices()
```

Listing 4.14: An example of the sensitivity analysis via Sobol indices for a PCE-based metamodel

Accuracy check The accuracy of the moment and sensitivity analyses greatly depends on the surrogate model's accuracy. There are two ways to inspect the accuracy of a trained metamodel in *BayesValidRox*. Using the `valid_metamodel` method of the `PostProcessing` module, one can perform a visual inspection by plotting the surrogate predictions against the simulation runs, either for specific samples, i.e., parameter sets or for randomly selected samples taken from the input marginals. For the former, the samples need to be passed to the method using the `samples` argument, and if available, the model outputs in a dictionary format using `model_out_dict`. The number of samples for the latter can be prescribed by `n_samples` argument.

The second option is to quantify accuracy with root-mean-square and validation error measures. The `PostProcessing` module offers the `check_accuracy` method for this purpose. It receives the samples (as a `numpy` array of shape `(n_samples,n_params)`) and the output as a dictionary containing the simulation outputs for all the model outputs. Alternatively, one can only define the number of randomly drawn samples from the input parameter space defined by the `Input` object.

```
# Plot to check validation visually.
PostPCE.valid_metamodel(n_samples=3)

# Compute and print RMSE error
PostPCE.check_accuracy(n_samples=300)
```

Listing 4.15: An example of the accuracy check for a metamodel

Note that, as for other cases, the `valid_metamodel` method generates plots containing the surrogate prediction against the corresponding simulation runs and stores them in the `Outputs_PostProcessing_calib` file. However, the `valid_metamodel` method only displays the errors in the console.

If the sequential method is selected as the experimental design method, one can plot the changes of the BME, KLD, Modified LOOCV, and the validation error (if requested) through the sequential learning iterations using the `plot_seq_design_diagnostics` method.

Regression quality check The success of a regression model depends on some fundamental assumptions about the underlying data (simulations) that it tries to model. Therefore, it is vital to check the quality of the regression model by verifying if the assumptions were reasonably satisfied. The key assumptions that need scrutinization are independence, homoscedasticity, and normality. The regression Listing 4.16 shows how to perform this regression quality check via *BayesValidRox*.

```
# Check the quality of your regression model
PostPCE.check_reg_quality()
```

Listing 4.16: An example of the regression quality check

BayesValidRox's post-processing plot residuals versus predicting variables allow for checking the independence assumption. The assumption holds if the residuals are distributed uniformly around the zero horizontal axes and do not form particular clusters.

To check the homoscedasticity assumption, *BayesValidRox* plots the fitted response values against the residuals. The assumption of homoscedasticity can be visually inspected by looking at the variance change. If the residuals' variance increases with the response values' magnitude, the problem does not respect homoscedasticity. The assumption of the normality of the data-generating process can be examined using the histogram and the quantile-quantile (Q-Q) plot of the normalized residuals. The latter is a graphical method for comparing the distribution of the normalized residuals with a normal distribution by plotting their quantiles against each other. The points in the Q–Q plot matching approximately the line with 45 degrees indicate that the residuals follow a normal distribution.

4.4 Surrogate-assisted calibration

BayesValidRox allows for a Bayesian parameter calibration using a surrogate model when dealing with computationally demanding models. The model calibration can be performed either as a part of a model validation workflow (Section 2.2) or only as a standalone calibration to obtain the posterior parameter distributions. The model calibration with *BayesValidRox* can be performed via a relatively straightforward rejection sampling method or a more sophisticated method, MCMC. For more information on these Bayesian inference methods, see Section 2.1.3.

As before, an instance of the `BayesInference` must be created first. It receives the already created metamodel object as the argument, as shown in line 3 in Listing 4.17. For selection of the rejection sampling inference method, one only passes the string `'rejection'` to the `inference_method` attribute. However, when an MCMC based inference is preferred, one needs to set the `inference_method` to `'MCMC'`. Moreover, some MCMC-specific parameters must be prescribed as a dictionary to the `mcmc_params` attribute. The complete list of these MCMC parameters along with their default values is presented in Listing 4.17 in lines 14 to 22. Note that the user-defined values overwrite the default values. Listing 4.17 shows two available inference options with the required specifications.

```
1  from bayesvalidrox import BayesInference, Discrepancy
2
3  BayesOpts = BayesInference(MetaModel)
4  BayesOpts.emulator = True # Default
5  BayesOpts.plot_post_pred = True # Plot the posterior predictives
6
7  # Select the inference method
8  # ----- Option I: Rejection Sampling -------
9  BayesOpts.inference_method = 'rejection' # Default
10
11 # ----- Option II: MCMC -------
12 BayesOpts.inference_method = 'MCMC'
13 # Set the MCMC parameters passed to self.mcmc_params
14 BayesOpts.mcmc_params = {
15     'init_samples': None,  # initial samples
16     'n_walkers': 100,  # Number of walkers (chains)
17     'n_steps': 100000, # Number of maximum steps per walker
18     'n_burn': 200,     # Number of burn-in steps
19     'moves': None,     # Moves for the emcee sampler
20     'multiprocessing': False, # Wether to use multiprocessing or not
21     'verbose': False   # Verbosity
22     }
23
24 # ----- Define the discrepancy model -------
25 # Known discrepancy
26 DiscrepancyOpts = Discrepancy('')
27 DiscrepancyOpts.type = 'Gaussian'
28 known_sigma_squared = {
29     'Z': np.array([1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0])
```

```
}
DiscrepancyOpts.parameters = known_sigma_squared # can be dict or
    pandas dataframe
BayesOpts.Discrepancy = DiscrepancyOpts

# Start the calibration/inference
Bayes = BayesOpts.create_inference()
```

Listing 4.17: An example of the surrogate-assisted Bayesian inference with a metamodel

When `None` is passed to the MCMC parameter `moves`, the `emcee.StrechMove` method for sampling proposals is selected. The list of the supported moves can be found on the *emcee* website[†]. Setting `verbose` to `True` prints the convergence criteria in the sampling process every 50 steps and plots the chains' evolution in the sampling process. This plot alongside the posterior distribution plot and the *emcee* backend file are stored in a `Outputs_Bayes_ModelName_Calib` directory. Moreover, prior (in case of the rejection sampling) and posterior predictive plots with their corresponding values in `hdf5` formats can also be found in the same directory.

Discrepancy model The model discrepancy, represented by ϵ in Equation (2.5), includes the effects of the measurement, model inaccuracy, and other sources of errors. In Listing 4.17, the discrepancy model has been defined to follow a multivariate Gaussian normal distribution with zero mean and covariance with diagonal entries defined as `known_sigma_squared`.

In some cases, the errors are not known or only partly known *a priori*. *BayesValidRox* offers the possibility of jointly inferring the discrepancy for each model output. To do so, one is expected to pass a discrepancy model containing the known and/or the unknown part as shown in Listing 4.18. Then, the discrepancy object model needs to be given to `BayesOpts.Discrepancy` attribute.

```
# ----- Define the discrepancy model -------
# Known discrepancy
known_DiscrepancyOpts = Discrepancy('')
known_DiscrepancyOpts.type = 'Gaussian'
known_sigma_squared = {
    'Z': np.array([1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0])
```

[†]https://emcee.readthedocs.io/en/stable/user/moves/

```
    }
known_DiscrepancyOpts.parameters = known_sigma_squared

# Unknown discrepancy -> to be jointly inferred with input parameters
DiscOutputOpts = Input()
DiscOutputOpts.add_marginals()
DiscOutputOpts.Marginals[0].name = '$\sigma^2_{\epsilon}$'
DiscOutputOpts.Marginals[0].dist_type = 'uniform'
DiscOutputOpts.Marginals[0].parameters =  [0, 4]
unknown_DiscOutputOpts = Discrepancy(DiscOutputOpts)

# Fully Unknown discrepancy
BayesOpts.Discrepancy = unknown_DiscOutputOpts

# Partially Unknown discrepancy
BayesOpts.Discrepancy = {
    'known': known_DiscrepancyOpts,
    'infer': unknown_DiscOutputOpts
    }
```

Listing 4.18: An example of the discrepancy model definition for surrogate-assisted Bayesian inference

4.5 Validation / Model comparison

As mentioned in Section 2.2, the updated information on the distribution of the input parameters resulting from the calibration process can be used for model validation. In the validation step, models are compared with a different measurement data set from what has been used to calibrate the model. In a multimodel validation case, a model inter-comparison step is also desired (see Section 2.2). This approach allows for a rigorous validation benchmarking that compares the model based on their predictive capabilities and also quantifies their similarities. For single model validation, the computed BME can be compared with the theoretical upper limit TOM, described in Section 2.2.2. One can state how likely the model is valid by comparing these two BME distributions – e.g., using statistical significance measures, such as p-value.

4.5.1 Single model validation via TOM

For a single validation, the module `BayesInference` and `Discrepancy` are required. The former starts the model validation process, while the latter defined the discrepancy term, as discussed earlier. We can account for the uncertainties in the validation metric BME assuming noisy data. This is done by bootstrapping.

```
from bayesvalidrox import BayesInference, Discrepancy

BayesOpts = BayesInference(MetaModel)
BayesOpts.emulator = True # Default

# Bootstrap for BME and BF
BayesOptsValid.bootstrap = True
BayesOptsValid.n_bootstrap_itrs = 10
BayesOptsValid.bootstrap_noise = 0.05

# ----- Define the discrepancy model -------
# Known discrepancy
DiscrepancyOpts = Discrepancy('')
DiscrepancyOpts.type = 'Gaussian'
known_sigma_squared = {
    'Z': np.array([1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0])
}
DiscrepancyOpts.parameters = known_sigma_squared # can be dict or
   pandas dataframe
BayesOpts.Discrepancy = DiscrepancyOpts

# Start the validation
Bayes = BayesOpts.create_inference()
```

Listing 4.19: An example of the surrogate-assisted Bayesian single-model validation

Setting `bootstrap` to `True` indicates that the bootstrapped Bayesian analysis is performed to assess the model validation. The plot containing the BME distribution of the model against that of the TOM can be found in the directory `Outputs_Bayes_{ModelName}_Valid`.

4.5.2 Multimodel comparison via justifiability analysis

After training the surrogate models for all the competing computational models, the model comparison can be started. It requires the model objects, stored in a dictionary and a dictionary containing the required options. Moreover, a discrepancy object is required similar to the Bayesian calibration. This can be a dictionary containing different `Discrepancy` objects for different models or one common `Discrepancy` object for all models.

```
# ----- Discrepancy model -------
# Known discrepancy
known_DiscrepancyOpts = Discrepancy('')
known_DiscrepancyOpts.type = 'Gaussian'
known_sigma_squared = {
    'Z': np.array([1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0])
    }
known_DiscrepancyOpts.parameters = known_sigma_squared

# ----- Multimodel comparison -------
from bayesvalidrox import BayesModelComparison
# Instantiate the object
BayesOpts = BayesModelComparison(
    justifiability=True, # Justifiability analysis
    n_bootstarp=1000, # Number of bootstrap iteration
    perturbed_data=perturbed_data, # pass perturbed observation data
    just_n_meas=2 # Confusion matrix for every 2 measurement points
    )

# Define the metamodel dictionary
meta_models = {
    "linear": L2_MetaModel,
    "exponential": NL4_MetaModel,
    "cosine": NL4_MetaModel
    }

# Define the option dictionary
opts_bootstrap = {
    "n_samples": 10000, # Number of parameter samples
    "n_bootstrap_itr": 1000, # number of bootstraping
    "Discrepancy": known_DiscrepancyOpts, # Discrepancy object
    "emulator": True, # Use emulator or the original model
```

```
    "plot_post_pred": True # Plot the posterior predictives
    }

# Start the model comparison
output_dict = BayesOpts.create_model_comparison(
    meta_models,
    opts_bootstrap
    )
```

Listing 4.20: An example of the surrogate-assisted multimodel comparison

If the `justifiability` parameter for `BayesModelComparison` is set to `False`, the module will only yield the multi-model comparison using model weights (Section 2.2.2) and Bayes factor (Section 2.2.1). Otherwise, a justifiability analysis (Section 2.2.2) is performed, in addition to the mentioned analyses. `output_dict` includes the Bayes factor and the model weights dictionaries for the BMS and justifiability analysis. The corresponding plots are stored in a directory named `Outputs_Comparison`.

The current chapter introduced *BayesValidRox*, an open-source, object-oriented package. This serves as a user guide to perform an automated surrogate-based sensitivity analysis, Bayesian calibration-validation and multi-model comparison. The next three chapters are dedicated to the application of the proposed framework in the last chapters to models describing flow and transport in porous media.

5 Application I: Flow Simulation Models in Fractured Porous Media

Flow in porous media is often characterized by very strong heterogeneities, particularly fractures, whose influence is essential for understanding the overall systems' behavior in many natural and technical applications. This chapter presents the application of the proposed surrogate-based Bayesian multi-model comparison framework to perform an uncertainty-aware comparison of two models for flow simulation in fractured porous media. This study was a joint effort of the researchers in Collaborative Research Center 1313 of the University of Stuttgart as a project in the 1st Summer School in 2019.

5.1 Introduction

Fractures can be found in many porous media such as glaciers, soils, rocks, woods, and concrete. They provide conduits or barriers for fluid flow that can alter flow and transport behavior in these media. For example, the flow pattern can be dominated mainly by the flow through connected channels formed by fractures. Flow in fractured porous media research has gained attention in the last decade to investigate the subsurface processes and has been used in biological applications and material science. According to Berre et al. [2019], the development of the high-resolution models primarily revolves around three issues: a) accurate representation of flow inside the fracture, b) good flow representation in the interaction between fractures and the neighboring porous media, and c) ability to handle complex fracture network structures.

Many approaches for modeling flow processes in fractured porous media have been developed over the last decades. The modeling challenges include the interaction between structural properties of fractures and fracture networks with occurring dynamic

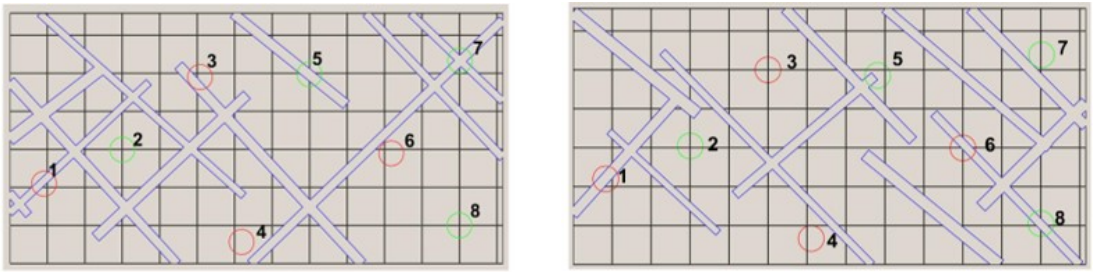
processes in the domain. The strong discontinuities of fluid velocities induced by the fractures cannot be adequately described in averaged descriptions. Furthermore, another challenge is to account for the alteration of the fracture networks resulting from the flow-related mechanical and chemical fluid-solid interactions. Berre et al. [2019] address these challenges and provide an overview of the standard conceptual models and discretization approaches with an emphasis on the dominating effects of fractures on flow processes. For more information about the concepts and mathematical models, the reader is referred to Berkowitz [2002] and Neuman [2005].

5.2 Problem description

The available models to represent the flow in fractured porous media vary drastically in their geometric details, simplifications, and abstraction level. In this study, two models were analyzed: the first model (B01) employs a phase-field representation with a flow model inside the fractures. In contrast, the second model (B03) uses a sharp fracture representation. As a joint model base, a one-phase Darcy flow was considered in the porous medium with dominant fractures, resulting in a pressure distribution throughout the porous medium. The pressure distribution across the fractures is treated slightly differently in the considered models, but all aim to fit the physical observations. The pressure values were measured on a manufactured fractured porous media block adjusted to the simulation benchmark. The results of the simulations have been compared with experimental data to validate the solvers and make a model comparison. In what follows, I introduce the experimental setup and the numerical models.

5.2.1 Experimental setup

The experiments considered two different cases with connected and disconnected fractures. In the first case, the fractures are connected from one side of the sample to the other, as shown in Figure 5.1a, while they are disconnected in the second case (Figure 5.1b). Additionally, measuring pressure values were obtained at the inlet, outlet and eight intermediate points of the disconnected/connected fracture network sample. This was accomplished with six tested and calibrated sensors (four at the intermediate points and two at in-/outlet locations).



(a) Setup with connected fractures (b) Setup with disconnected fractures

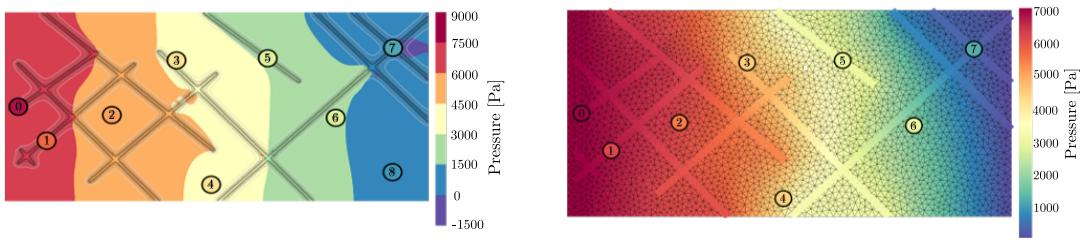
Figure 5.1: Two experimental setups: (a) connected and (b) disconnected fracture networks.

Due to the lack of space on the sample for installing sensors at all 8 measurement points, the pressure values were measured separately on the sample, four sensors at the time, for different flow rates. The fluid used in this experiment was distilled water. The fluid's flow rate was carefully adopted in the range of 0.1 to 0.5 [ml/min] to avoid inducing critical deformation into the sample, which may cause pressure diffusion.

5.2.2 Conceptual models

As mentioned earlier, I investigate two model variants for flow simulation in fractured porous media in this study. The first model (B01) employs a phase-field fracture representation [Kuhn and Müller, 2010] and simulates the solid-fluid interaction within the porous medium under the consideration of Biot's theory of consolidation [Verruijt, 2010]. Moreover, Darcy's law describes the fluid flow in the porous medium where an additional permeability tensor increases the permeability within the cracks. It models Poiseuille-type flow within the cracks and is derived from the lubrication theory. The solution is obtained by the finite element method in *FEAP*[Taylor, 2014].

Model B03 takes a discrete fracture network approach. All fractures are geometrically resolved to conform to the mesh. A mixed-dimensional model is used for Darcy flow both in the bulk porous medium and the fractures. The permeability of the fractures is determined by the Poiseuille approximation using the hydraulic diameter. The solution is obtained by a Finite-Volume method implemented in DuMu^{x} [Koch et al., 2021a].



(a) Model B01 (b) Model B03

Figure 5.2: The model domains of the connected case for (a) Model B01 and (b) Model B03

5.2.3 Errors and uncertainties

Bayesian statistics updates the prior belief on the model by comparing the model responses with the measured data. It can include not only the errors and uncertainties in the observed data, but it can also take into consideration other sources of errors, such as numerical errors and surrogate model's error. These errors can be considered in the covariance matrix $\mathbf{\Sigma}$ in Equation (2.6), assuming that they follow a normal distribution. In the following, the parameter uncertainty and three sources of error used in the calibration and validation benchmark of the flow simulation models are introduced.

Parameteric uncertainty Uncertain parameters and their ranges for the model validation analysis are summarized in Table 5.1. The sample's depth was considered in

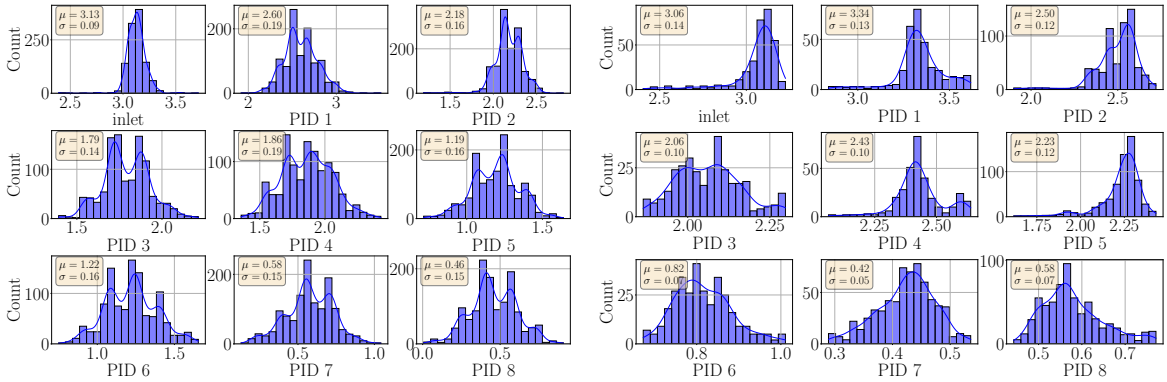
Table 5.1: List of considered uncertain parameters and their defined distributions for both models.

Parameter name	Range	Unit	Distribution type
Permeability (porous medium), k	$[10^{-13}, 10^{-11}]$	m/s	uniform
Sample's depth, d	$[8 \cdot 10^{-5}, 10^{-4}]$	m	uniform
Outflow pressure, P_{out}	$[400, 2000]$	Pa	uniform

the flow calculations at the boundary condition.

Measurement errors In consultation with the experimentalists, I used the standard deviation of the measured value for each measurement sensor as the experimental error. These values are extracted separately for two cases and different flow rates for

calibration and validation. The pressure measurement's distribution for two different cases for the calibration step for the connected and disconnected case, respectively, are shown in Figure 5.3.



(a) Connected case ($Q = 0.2\,ml/min$) (b) Disconnected case ($Q = 0.1\,ml/min$)

Figure 5.3: The pressure distributions of nine sensors used for the calibration for (a) the connected and (b) the disconnected case

Numerical errors Here, I only considered the discretization error that originates from a certain choice of meshing size. Following Oberkampf and Roy [2010], I took a heuristic approach to quantify this error in which I used a generalized Richardson extrapolation based on different mesh spacing to estimate the error. The Richardson extrapolation takes the following form:

$$f_k = \bar{f} + g_p h_k^{\hat{p}} + \mathcal{O}\left(h_k^{\hat{p}+1}\right). \tag{5.1}$$

f_k denotes the solution to the discrete equation on a mesh with spacing h_k (known), $\bar{f}$ stands for the exact solution to the original PDE. g_p is the error term coefficient and $\hat{p}$ indicates the observed order of accuracy. Here, we seek the error with an order of one. The unknowns, $\bar{f}$ and g_p, can be easily determined via a least square method using the solutions obtained for different meshing sizes.

Surrogate model's errors As discussed in previous chapters, replacing the computational models with surrogates may introduce additional errors to the inference process. To include this error, we test the surrogate models trained with 20 simulation runs (test sets) which are different from the training sets consisting of 180 runs. Comparing the surrogates' prediction with the results from the test sets, we observed an acceptable

prediction accuracy with the validation errors (Section 3.1.3) between 10^{-5} and 10^{-9} for all models. Moreover, we estimated Mean Square Error (MSE) for each surrogate model, which is a good estimate of the surrogate error variance [Xu and Valocchi, 2015]. When evaluating the likelihood $p(\mathcal{Y}|\mathcal{M}_k, \boldsymbol{\theta}_k)$ in Equation (2.6), I added a diagonal matrix Σ_{PCE} with elements $\sigma^2_{\text{PCE},i} = MSE_i$, $i = 1, 2, ..., N_{\text{out}}$ to $\boldsymbol{\Sigma}$, assuming that the surrogate errors are independent and follow a normal distribution with zero mean.

5.2.4 Solution procedure

The following steps are taken for this model comparison study within the Bayesian approach. First, a surrogate model is trained based on the simulation results obtained by the computational models based on the pressure readings at 18 sensors (nine sensors for connected and nine for the disconnected case) as shown in Figure 5.1. Second, the Bayesian updating is performed for each model, in that the prior knowledge on the distribution of uncertain parameters is updated by comparing the model outputs with the measurement, using Equation (2.4). This step was performed with the MCMC method (Section 2.1.3) and yielded the so-called posterior distributions with narrower shapes. Afterward, new surrogate models were trained based on the newly obtained posterior distributions. Finally, samples (parameter sets) of the resulting posterior distribution from calibration were evaluated using the newly trained surrogates to obtain the posterior predictive distributions of each sensor. The validation metric BME was computed by comparing the posterior predictive distribution with a newly observed data set obtained by another flow rate. Moreover, following Schöniger et al. [2015a], we perturbed the reference data with some additive noise to account for uncertainty associated with the BME values, the resulting Bayes factors, and posterior model weights. With this approach, we investigated the impact of other possible ignored sources of errors on the validation metrics.

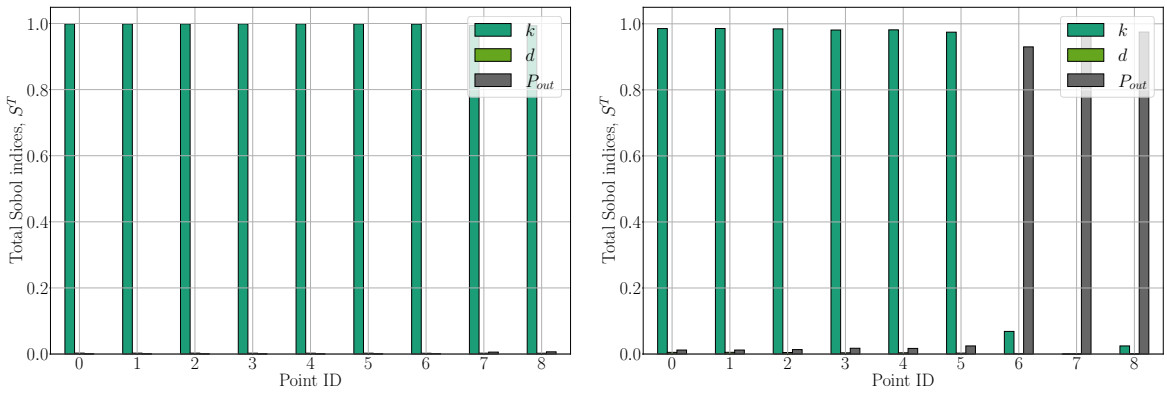
5.3 Results and discussions

The current section presents insights into the analysis of predictive abilities and model comparison using the procedure described earlier. Additionally, the results of global

sensitivity analyses are discussed in what follows. Moreover, the models' BME distributions are compared to the upper limit for model performance, TOM, as defined in Section 2.2.2.

5.3.1 Global sensitivity analysis

The surrogate modeling with PCE offers a property for global sensitivity analysis, as discussed in Section 3.1.7. This property requires a simple postprocessing step after training the surrogate models. In this section, I will analyze how the variability of the model response quantities (pressure readings) at the data extraction points is affected by the variability of each input variable or combinations thereof. This is achieved via sensitivity analysis using Sobol indices discussed in Section 3.1.7. Figures 5.4a and 5.4b display the total Sobol indices for model B01 for two considered cases. The total Sobol

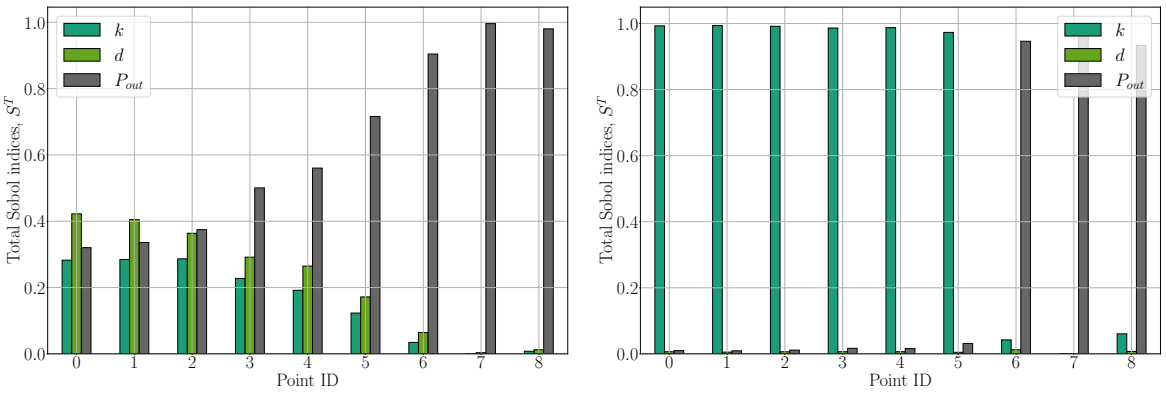


(a) Connected case (b) Disconnected case

Figure 5.4: The Sobol indices of the model B01.

indices for model B03 are illustrated in Figures 5.5a and 5.5b.

The Sobol indices of model B01 reveal that permeability (k) has the highest influence on the pressure variability in the connected case. However, the outflow pressure (P_{out}) shows higher sensitivity for the sensors close to the right boundary for the disconnected case. The sample depth (d) has no impact on the pressure values. In the case of model B03, the effect of pressure output (P_{out}) on the results increases, moving from the inlet to the outlet of the domain. At the same time, the sample depth (d) and the permeability (k) show approximately similar impacts and decrease from left to



(a) Connected case (b) Disconnected case

Figure 5.5: The Sobol indices of the model B03.

right. Comparing Figures 5.4b and 5.5b makes it clear that the parameters of the two considered models exhibit the same sensitivity behaviors.

5.3.2 Analysis of the predictive capabilities

One of the standard practices for model validation is the visual comparison of the SRQs against the observed data. An uncertainty-aware validation via a Bayesian framework yields uncertainties associated with the model predictions. These uncertainties are related to the posterior predictive pressure distribution for each sensor. In this section, I will present the result of the analysis of the predictive ability of two discussed models by showing the parameters' updated (posterior) distribution after calibration obtained by the MCMC sampler for all models. Afterward, figures containing the posterior predictive of models versus the measured data are provided.

Model B01 Figure 5.6 presents the posterior distribution obtained via the Bayesian inference using the calibration (mean values in Figure 5.3). The 50 percent quantiles, alongside the 15 and 85 percent quantiles, are displayed on top of the histograms shown in the diagonal plots. The posterior distributions of the permeability k, and the outflow pressure P_{out} follow Gaussian distributions. The distribution of the sample height d is more concentrated around the values at the lower boundary.

The posterior predictive distribution can be simply obtained by evaluating the surrogate model trained after calibration for samples drawn from the posterior distributions

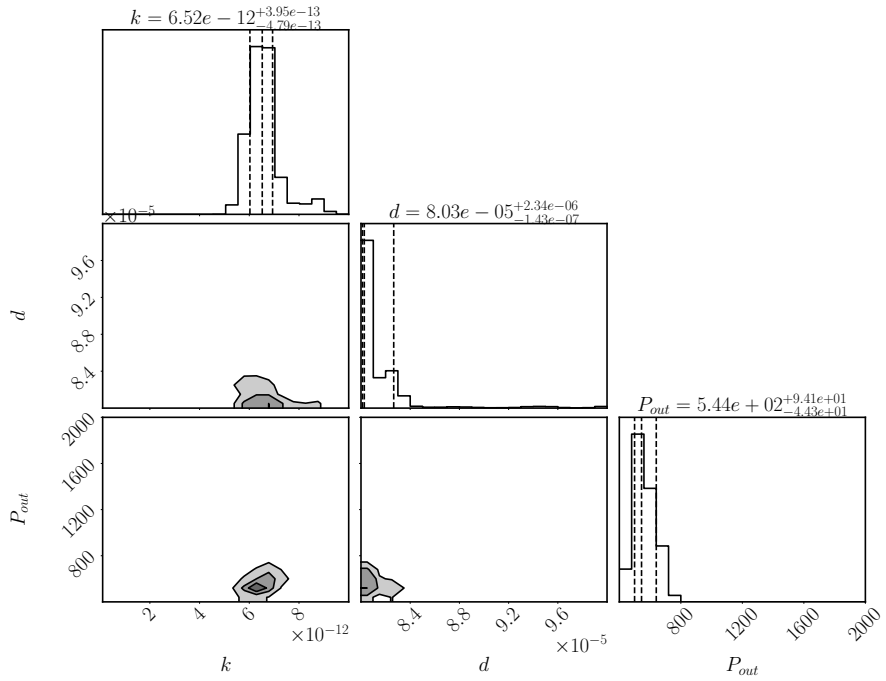
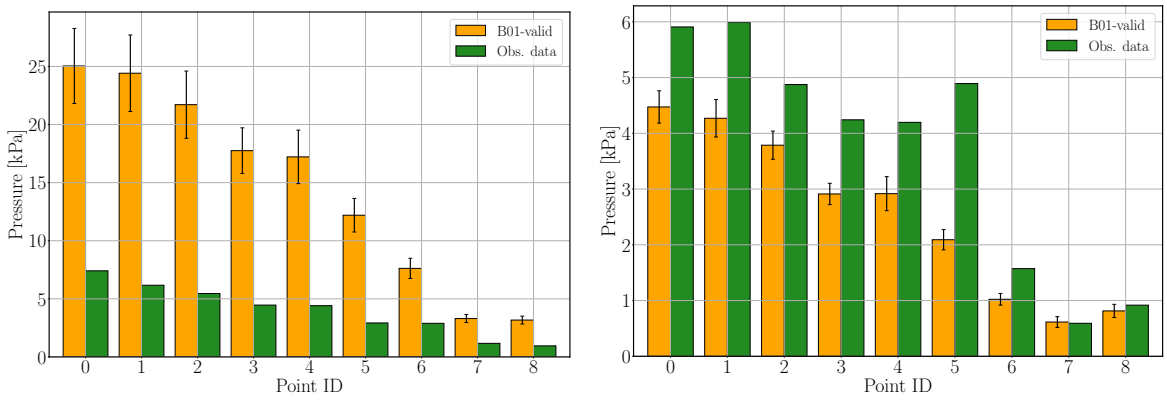


Figure 5.6: The posterior parameter distributions of model B01.

in Figure 5.6. I will present these distributions using a bar plot, with the bar showing the mean and 95 % credible interval marked with error bars. Figure 5.7 illustrates the posterior predictive plots of model B01 for the connected and disconnected cases. Model B01's pressure predictions show substantial deviation from the observed data



(a) Connected case (b) Disconnected case

Figure 5.7: The posterior predictive plots of the model B01.

in the experiment, especially for the connected case. This so-called model error was also observed in the calibration phase. Further investigation of the moments and stan-

dard deviation and the training model evaluations confirmed this hypothesis. As for the pressure values in the disconnected case, the simulated pressure values slightly underestimate the observed data.

Model B03 The posterior distribution of model B03 after the calibration is presented in Figure 5.8. The posterior distribution of the permeability k follows a Gaussian

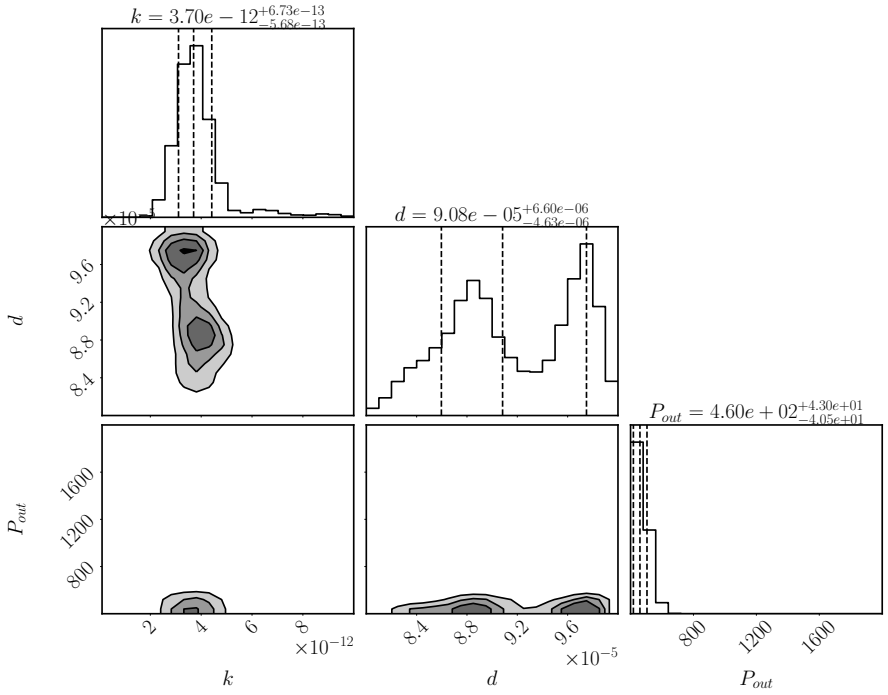
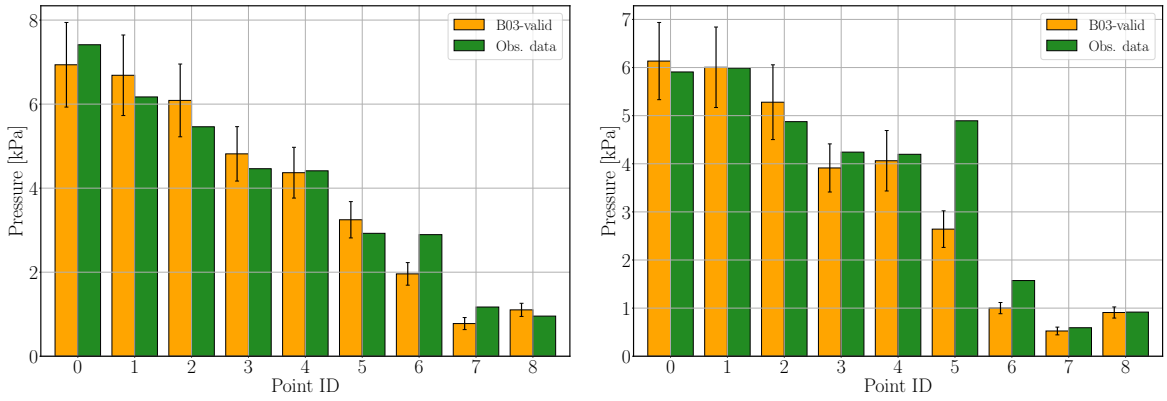


Figure 5.8: The Posterior parameter distribution of model B03.

distribution. However, the calibration resulted in a multi-modal distribution for the sample height d. The distribution of the outflow pressure P_{out} is more concentrated in the values to the lower boundary, suggesting that a prior distribution with a broader range can be used.

Figure 5.9 shows the posterior predictive plots of model B03 for a) the connected and b) disconnected cases. The visual inspection of the diagrams in Figure 5.9 indicates that model B03 better matches the validation data. Except for sensors 6 and 7 for the connected case and 5 and 6 for the disconnected case, all observed pressure values fall in the posterior predictive credible intervals.



(a) Connected case (b) Disconnected case

Figure 5.9: The posterior predictive plots of the model B03.

5.3.3 Model validation and comparison

This subsection presents the results of the Bayesian model validation and multi-model comparison. I performed a model comparison employing the so-called posterior model weights according to the BMS explained in Section 4.5. Such an analysis offers an aggregated comparison of a model's outputs to the validation data set. The use of the advanced surrogate representation allows the assessment of the uncertainty of the BME values and the corresponding model weights. Table 5.2 presents a detailed statistical summary of the posterior model weights and offers a ranking. It also reports the information regarding the post-calibration uncertainty with the help of the deviation regarding 25% and 75% percentiles.

Table 5.2: Posterior model weights after validation.

Model	Model weights	Rank
B01	$0.001^{+0.000}_{-0.000}$	2
B03	$0.999^{+0.000}_{-0.000}$	1

The expected model weights under noisy data assumption convey a clear model ranking favoring model B03. This clear ranking is in accordance with the conclusions drawn from a visual inspection of the posterior predictive plots in Figures 5.7 and 5.9. The considerable pressure differences between the simulated and the measured data for the connected case via model B01 led to smaller BME values than model B01. Consequently, the posterior weights of model B03 obtained a value close to the maximum

weight possible.

As discussed earlier, Bayesian statistics offers BME as a quantitative metric to assess the model performance. The model's validity can be examined by comparing its BME with that of a theoretically optimal model. Figure 5.10 shows the distribution of the

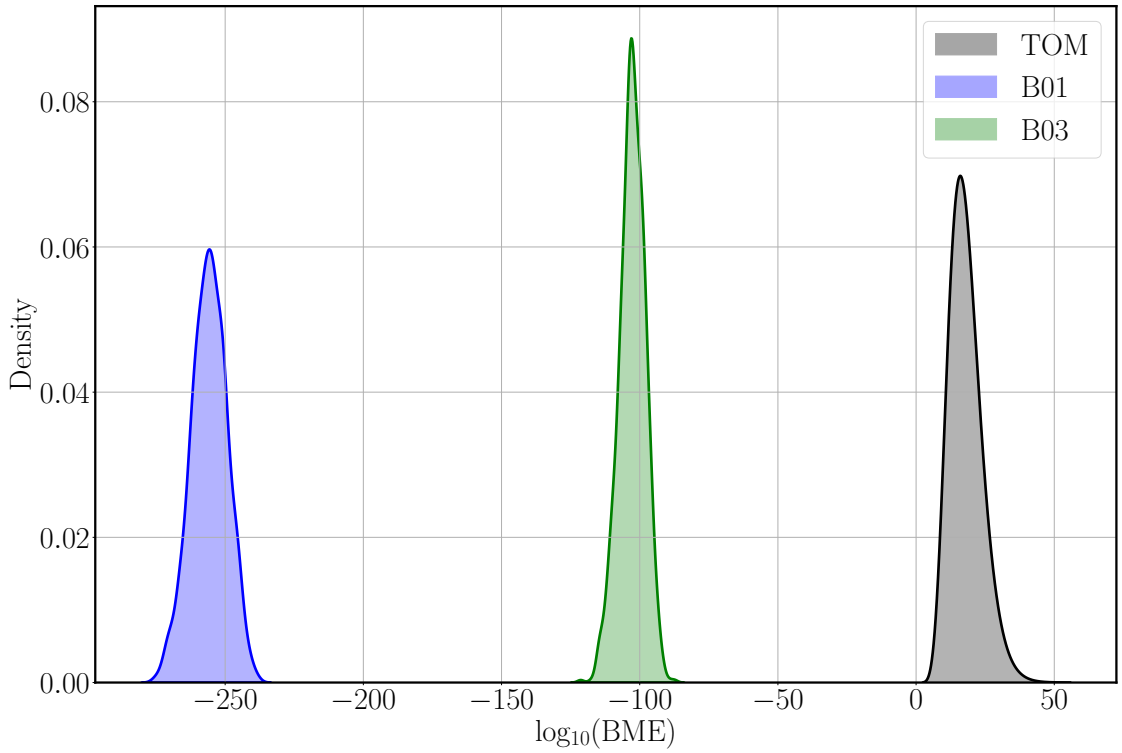


Figure 5.10: The BME distributions for the competing models with the perturbed data set

BME for the models B01 and B03, as well as the upper limit performance. These values are obtained by varying the measurement data set to account for the uncertainty associated with BME. Model B03 clearly outperforms B01, as its BME distribution is closer to that of TOM. However, it cannot be considered valid since its BME values do not overlap the upper limit provided by TOM. This discrepancy can be attributed to the discrepancies of simulated and measured pressure values in the posterior predictive plots at some sensors in Figure 5.9.

One approach to further compare models is hypothesis testing in a Bayesian setting. The competing models are compared pairwise in the BHT based on the so-called Bayes factors, which comprise the models' BME values ratio. The Bayes factor is a measure

of significance and quantifies the evidence of the hypothesis that one model is the data-generating model against a null hypothesis. The null hypothesis can be defined as one model being the best model within the model pair under investigation (Section 2.2.1). In that regard, one can consider TOM as an additional model and formulate hypotheses so that each model is compared with its counterpart and the TOM. Figure 5.11 presents the probability density functions of $\log_{10}(\mathrm{BF})$ using the perturbed pressure data sets in a three-by-three matrix. The significance levels in a $\log_{10}$-scale, introduced in Jeffreys [1961] are marked with vertical lines. Due to the large values for BFs, these vertical lines have approximately overlapped.

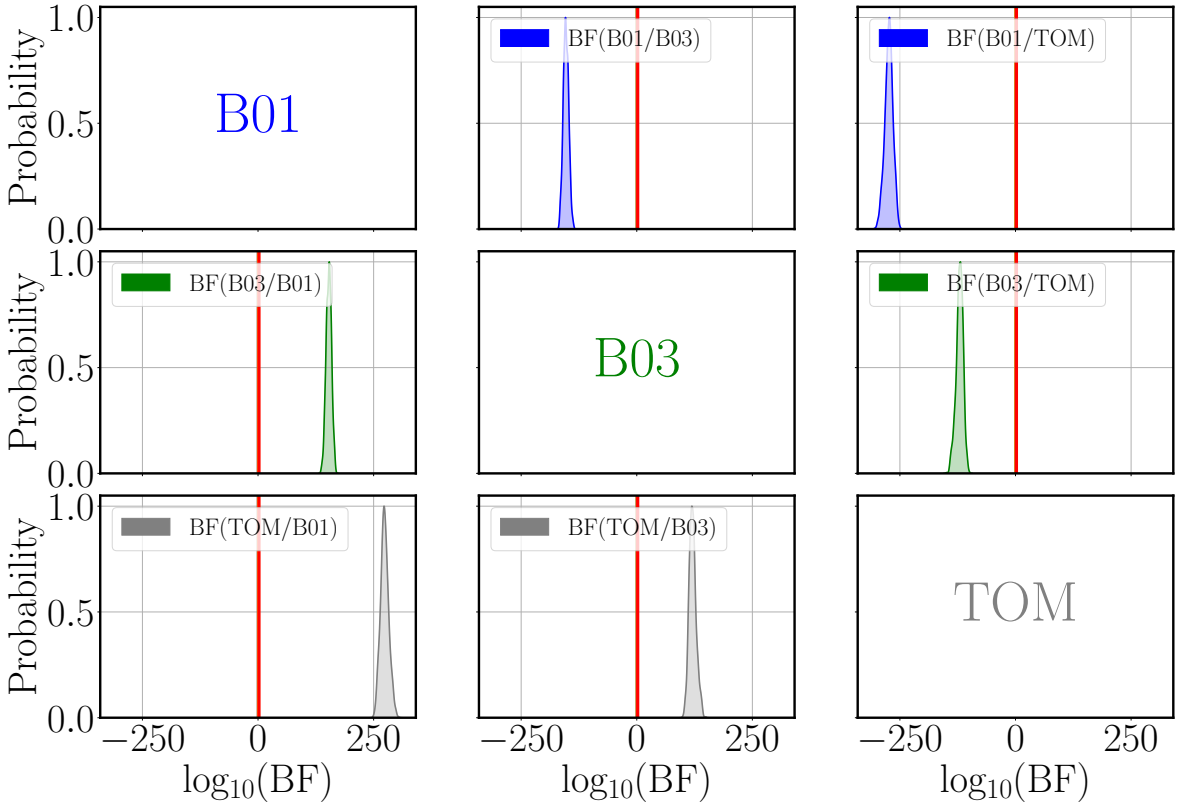


Figure 5.11: The pairwise comparison of models via BHT.

As expected from the model ranking, there is decisive evidence in favor of the model B03 against B01. This conclusion is based on the fact that the distribution of Bayes factors is located on the right-hand side of the significance levels in the plot in the first column and the second row. The last row indicates that both models are far from the optimal model described by TOM. This finding does not necessarily suggest further improvement of both models, but the measurement procedure and the experimental

results require revision. These considerations were not done due to the limited time of the summer school.

5.4 Summary and conclusions

In this chapter, I have applied the surrogate-assisted uncertainty-aware Bayesian validation framework to compare two computational models describing the fluid flow in fractured porous media. The models were compared in light of their abilities to reproduce the measured data. The parametric and other sources of errors were also addressed in the computations. In the end, the performance of the models was analyzed via a quantitative validation metric. The considered computational models were substituted by their PCE-based surrogate models to accelerate the uncertainty propagation in Bayesian calibration and validation. A benefit of employing these surrogates is that a sensitivity analysis without additional costs can be performed. This analysis is achieved using the so-called Sobol indices that are derived analytically from the expansion coefficients. The results of the analysis indicate that both models require improvement. Model B03 showed better performance suggested by both visual inspections of the posterior predictive plots and quantitative comparisons. However, the posterior predictive plots for this model revealed that the model predictions did not match the experimental results at some pressure sensors. This was the only reason that the distribution of the BME (validation metric) did not reach the achievable upper limit provided by TOM.

6 Application II: Bayesian Comparison of Conceptually Simplified Models to a Detailed Reference Model: Application to Coupling Free Flow and Porous-Medium Flow

Coupled free-flow and porous-medium systems play a significant role in many industrial, environmental, and biological settings, such as fuel cells, water flows in karst aquifers, blood flows in vessels, and living tissues. Flow interaction between the free-flow region and the porous-medium domain is highly involved and strongly interface-driven. The correct choice of interface conditions and proper model parameters for these systems is vital for physically consistent modeling and accurate numerical simulations of applications.

This chapter deals with the application of the surrogate-assisted Bayesian validation framework to coupling free flow and porous-medium flow. We studied the coupled flow problems' behaviors considering some benchmark cases, where a pore-scale resolved model provides the reference solution. We quantified the sensitivity of the models' parameters and their uncertainties after calibration. Moreover, the performance of models against the reference solution was compared within a Bayesian hypothesis testing framework. In what follows, I present the result of this benchmark study via a surrogate-assisted uncertainty-aware Bayesian validation framework. This study has been submitted to the Journal of Computational Geosciences [Mohammadi et al., 2022].

6.1 Introduction

A physically consistent description of flow processes in the entire coupled system, especially near the interface, is crucial for a better grasp of the interface-driven processes in applications such as solid-atmospheric interactions, drying processes during evaporation, and industrial filtration [Hanspal et al., 2009, Beaude et al., 2019, Jarauta et al., 2020]. A lot of effort has been made during the last decades in mathematical modeling and analysis of coupled flow systems. These mathematical models preserve mass, momentum, and energy conservation in both flow domains and across the fluid-porous interface. Navier-Stokes equations are generally applied to describe fluid flow in the free-flow domain, and multi-phase Darcy's law describes the fluid flow in porous media [Discacciati and Quarteroni, 2009, Eggenweiler and Rybak, 2020].

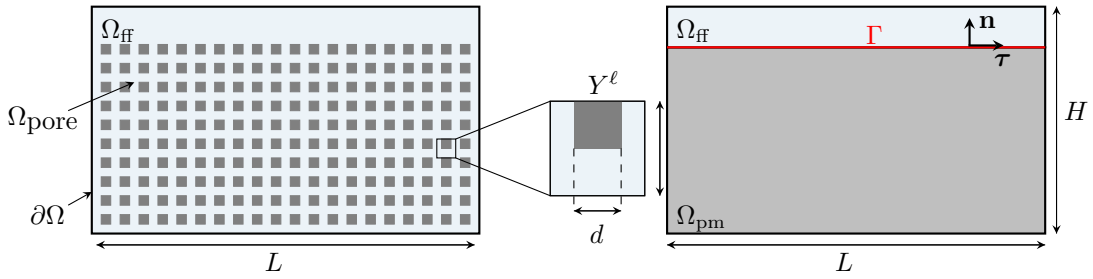
Based on the application of interest and the flow regime, this general system may be simplified in various ways. The most widely studied free-flow and porous-medium flow system is described by the coupled Stokes–Darcy equations with different sets of interface conditions, [e.g., Angot et al., 2017, Discacciati and Quarteroni, 2009, Goyeau et al., 2003, Jäger and Mikelić, 2009, Lācis and Bagheri, 2017, Ochoa-Tapia and Whitaker, 1995, Eggenweiler and Rybak, 2021]. The possibilities to conceptualize these coupling conditions could be regarded as conceptual uncertainty. This conceptual uncertainty is mainly related to the description of processes in the porous medium and near the interface, for which different mathematical models and coupling strategies are considered. Besides the conceptual uncertainty, each computational model contains parametric uncertainty, such as material parameters or interface location that must also be rigorously addressed.

A significant research challenge is to assess competing modeling concepts for coupling free flow and porous medium accurately and validate the corresponding computational models against an experiment or a reference solution in light of uncertainties. To this end, we analyzed three coupling concepts in a validation benchmark study using the reference data resulting from a detailed simulation. We used the literature's most widely studied coupled flow problem, namely the Stokes–Darcy problem. In this setting, the free-flow conceptualization is based on the Stokes equations in the free-flow domain for all discussed models. However, the way these models simulate the fluid flow in the porous medium and the set of coupling conditions imposed on the fluid–porous interface

varies. In what follows, the Stokes–Darcy problem and the mathematical description of the governing equations in the different system domains will be presented in Section 6.2.

6.2 Problem description

From a pore-scale perspective, we consider a two-dimensional flow domain Ω_{flow} consisting of the free-flow domain Ω_{ff} and the pore space Ω_{pore} of the porous medium. The porous-medium domain Ω_{pm} has a periodic structure composed by the repetition of the representative elementary volume (REV) (scaled unit cell) $Y^\ell = (0,\ell)\times(0,\ell)$, where ℓ is the microscopic length scale (Figure 6.1a). From a macroscopic point of view, the coupled domain $\Omega = \Omega_{\text{ff}} \cup \Omega_{\text{pm}}$ comprises the free-flow region Ω_{ff} and the porous-medium domain Ω_{pm}, separated by a sharp fluid–porous interface Γ (Figure 6.1b).



(a) Pore scale (b) REV scale

Figure 6.1: Geometrical setting at the considered scales.

We consider isothermal single-phase steady-state flow at low Reynolds numbers. The same fluid occupies the free-flow domain and fully saturates the porous medium. This fluid is supposed to be incompressible and to have constant viscosity. The porous medium is considered non-deformable, resulting in constant porosity.

6.2.1 Reference pore-scale resolved model

At the pore scale, fluid flow in the flow domain Ω_{flow} is governed by the Stokes equations,

$$\nabla\cdot\mathbf{v} = 0, \qquad -\nabla\cdot\mathbf{T}(\mathbf{v},p) = \mathbf{0} \qquad \text{in } \Omega_{\text{flow}}, \tag{6.1}$$

completed with the no-slip condition on the boundary of solid inclusions

$$\mathbf{v} = \mathbf{0} \quad \text{on} \ \ \partial\Omega_{\text{flow}} \setminus \partial\Omega, \tag{6.2}$$

and appropriate conditions on the external boundary $\partial\Omega$. Here, $\mathbf{v}$ and p denote the fluid velocity and pressure, $\mathbf{T}(\mathbf{v}, p) = \mu\nabla\mathbf{v} - p\mathbf{I}$ the stress tensor, $\mathbf{I}$ the identity tensor and μ the dynamic viscosity.

Resolving pore-scale information is computationally expensive for practical applications. Therefore, REV-scale model formulations, which accurately reflect the pore-scale flow processes, are often preferred and are investigated in this study. The pore-scale resolved model is used only as a reference for the model validation purposes. A finite-volume scheme on staggered grids, also known as MAC scheme [Harlow and Welch, 1965], is used to discretize the pore-scale model in Equations (6.1) and (6.2).

6.2.2 Subdomain models

In this study, we considered two different types of coupled models, for which the Stokes equations are used in the free-flow region Ω_{ff}. However, the porous domain Ω_{pm} is treated by different modeling concepts. The first type of model relies on the REV-scale description of the porous-medium domain using Darcy's law, whereas the second type of model follows a hybrid-dimensional approach, where a lower-dimensional pore-network model (PNM) is used to describe the fluid flow in the porous domain [Weishaupt et al., 2019, 2020].

Free-flow model As a common feature, both coupled models (REV-scale model, pore-network model) contain the incompressible, stationary Stokes equations for the description of fluid flow in the free-flow domain

$$\nabla\cdot\mathbf{v}_{\text{ff}} = 0, \quad -\nabla\cdot\mathbf{T}(\mathbf{v}_{\text{ff}}, p_{\text{ff}}) = \mathbf{0} \qquad \text{in} \ \ \Omega_{\text{ff}}, \tag{6.3}$$

where $\mathbf{v}_{\text{ff}}$ and p_{ff} represent the fluid velocity and pressure, respectively. Similar to the pore-scale model (Section 6.2.1), we employ MAC scheme for discretizing the Stokes equations in (6.3).

REV-scale porous-medium model At the REV-scale, fluid flow through the porous medium is described by the Darcy flow equations, which reads

$$\nabla \cdot \mathbf{v}_{\mathrm{pm}} = 0, \qquad \mathbf{v}_{\mathrm{pm}} = -\frac{\mathbf{K}}{\mu} \nabla p_{\mathrm{pm}} \quad \text{in } \Omega_{\mathrm{pm}}, \tag{6.4}$$

where $\mathbf{v}_{\mathrm{pm}}$ is the Darcy fluid velocity, p_{pm} is the fluid pressure, and $\mathbf{K}$ is the intrinsic permeability tensor, which is symmetric, positive definite, and bounded. Equations (6.4) are discretized with a vertex-centered finite-volume scheme, also known as box method [Hackbusch, 1989].

Pore-network porous-medium model Pore-network models [Blunt, 2017] consider a simplified yet equivalent representation of the porous geometry by separating the void space into larger pore bodies connected by narrow pore throats. Despite their low computational demand, a rather high degree of pore-scale accuracy can be achieved [Oostrom et al., 2016]. PNMs can also be combined with modeling approaches on different scales [Scheibe et al., 2015], such as Darcy-type continuum models [Balhoff et al., 2007, 2008, Mehmani and Balhoff, 2014] or free-flow models [Beyhaghi et al., 2016].

For the PNM, we require the conservation of mass for each pore body i (the intersection of two or more pore throats):

$$\sum_j Q_{ij} = 0, \qquad Q_{ij} = g_{ij}(p_i - p_j). \tag{6.5}$$

Here, Q_{ij} is the discrete volume flow rate in a throat connecting the pore bodies i and j, and the pressures defined at the centers of the pore bodies i and j are given by p_i and p_j (Figure 6.2). Equation (6.5) represents a finite-volume discretization scheme with a two-point flux approximation, see Weishaupt et al. [2020], Koch et al. [2021b] for further details. The pore throat geometry and fluid properties determine the total conductance g_{ij}. Considering the pressure losses both within the pore bodies and throats, we use

$$g_{ij} = \left(g_{t,ij}^{-1} + g_{p,i}^{-1} + g_{p,j}^{-1}\right)^{-1} , \tag{6.6}$$

where $g_{t,ij}$ is the conductance of a throat ij while $g_{p,i}$ and $g_{p,j}$ are the conductances of the adjacent pore-body halves (Figure 6.2). In Figure 6.2, throat ij connects the

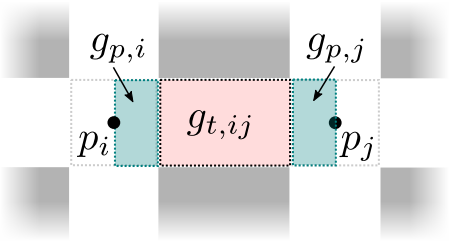


Figure 6.2: Schematic contribution to total conduction for the PNM.

pore bodies i and j at the centers of which the pressures p_i and p_j are defined. $g_{t,ij}$ is the throat conductance valid for the region marked in light red. $g_{p,i}$ and $g_{p,j}$ are the conductances defined for the pore body halves marked in teal.

Simple analytical expressions for g_{ij} are available in the literature [Patzek and Silin, 2001] for certain geometries. Usually, we determine g_{ij} via numerical upscaling [Mehmani and Tchelepi, 2017], whereas for this study, we consider it to be an additional uncertain parameter. In the following, we only refer to $g_{p,i}$, as for the given geometry $g_{p,i} = g_{p,j}$ for interior throats. At interface throats, one of the half-pore-body conductance is zero.

6.2.3 Coupling concepts

A variety of REV-scale coupling concepts for the Stokes–Darcy system (6.3)–(6.4) is available in the literature. In this study, we considered the most widely used set of interface conditions, based on the Beavers–Joseph condition and the recently developed generalized conditions [Eggenweiler and Rybak, 2021]. If the PNM (6.5) is used in the porous medium, separate coupling conditions, suitable for the pore-scale description of interface exchange processes, must be considered.

Classical coupling conditions (REV-scale model) The most commonly used interface conditions are the *conservation of mass*

$$\mathbf{v}_{\mathrm{ff}}\cdot\mathbf{n} = \mathbf{v}_{\mathrm{pm}}\cdot\mathbf{n} \qquad \text{on } \Gamma, \tag{6.7}$$

the *balance of normal forces*

$$-\mathbf{n}\cdot\mathbf{T}(\mathbf{v}_{\mathrm{ff}}, p_{\mathrm{ff}})\mathbf{n} = p_{\mathrm{pm}} \qquad \text{on } \Gamma, \tag{6.8}$$

and the *Beavers–Joseph condition* [Beavers and Joseph, 1967] for the tangential component of velocity

$$(\mathbf{v}_{\mathrm{ff}} - \mathbf{v}_{\mathrm{pm}})\cdot\boldsymbol{\tau} - \frac{\sqrt{\mathbf{K}}}{\alpha_{\mathrm{BJ}}}\boldsymbol{\tau}\cdot\nabla\mathbf{v}_{\mathrm{ff}}\mathbf{n} = 0 \qquad \text{on } \Gamma. \tag{6.9}$$

Here, $\alpha_{\mathrm{BJ}} > 0$ is the Beavers–Joseph parameter, $\mathbf{n}$ is the normal unit vector on Γ pointing outward from the porous medium, $\boldsymbol{\tau}$ is a tangential unit vector on Γ and $\sqrt{\mathbf{K}} = \sqrt{\boldsymbol{\tau}\cdot\mathbf{K}\boldsymbol{\tau}}$ (Figure 6.1).

The Beavers–Joseph interface condition (6.9) was postulated for flows parallel to the interface [Beavers and Joseph, 1967]. Eggenweiler and Rybak [2020] and Eggenweiler and Rybak [2021] show that this condition is not suitable for arbitrary flow directions to the porous bed. However, it is routinely applied in the literature to multidimensional flows [Discacciati and Gerardo-Giorda, 2018, Hanspal et al., 2009].

Generalized coupling conditions for arbitrary flows (REV-scale model) An alternative to the classical interface conditions for Stokes–Darcy problems are the generalized coupling conditions derived rigorously in Eggenweiler and Rybak [2021] via homogenization and boundary layer theory. These conditions are valid for arbitrary flow directions to the fluid–porous interface and read

$$\mathbf{v}_{\mathrm{ff}}\cdot\mathbf{n} = \mathbf{v}_{\mathrm{pm}}\cdot\mathbf{n} \qquad \text{on } \Gamma, \tag{6.10}$$

$$\begin{aligned} p_{\mathrm{pm}} &= -\mathbf{n}\cdot\mathbf{T}(\mathbf{v}_{\mathrm{ff}}, p_{\mathrm{ff}})\mathbf{n} \\ &\quad + \mu N_s^{\mathrm{bl}}\boldsymbol{\tau}\cdot\nabla\mathbf{v}_{\mathrm{ff}}\mathbf{n} \qquad \text{on } \Gamma, \end{aligned} \tag{6.11}$$

$$\begin{aligned} \mathbf{v}_{\mathrm{ff}}\cdot\boldsymbol{\tau} &= -\ell N_1^{\mathrm{bl}}\boldsymbol{\tau}\cdot\nabla\mathbf{v}_{\mathrm{ff}}\mathbf{n} \\ &\quad + \mu^{-1}\ell^2\sum_{j=1}^{2}\frac{\partial p_{\mathrm{pm}}}{\partial x_j}\mathbf{M}^{j,\mathrm{bl}}\cdot\boldsymbol{\tau} \qquad \text{on } \Gamma. \end{aligned} \tag{6.12}$$

Here, $\mathbf{M}^{j,\mathrm{bl}}$, N_1^{bl} and N_s^{bl} are boundary layer constants introduced in Eggenweiler and Rybak [2021]. For the generalized coupling conditions, the interface can be located at a distance $\mathcal{O}(\ell)$ from the top of the first row of solid inclusions, where ℓ denotes the characteristic pore size. Based on the chosen interface position and the pore geometry, the effective coefficients appearing in conditions (6.10)–(6.12) are computed numerically using the theory of homogenization and boundary layers [Carraro et al., 2015, Hornung,

1996, Jäger and Mikelić, 2000, 2009]. This is the main advantage of the generalized interface conditions, besides their suitability for arbitrary flows in coupled Stokes–Darcy systems.

Coupling conditions for the pore-network model Each intersection of a pore body i with the free-flow domain boundary yields a pore-local discrete interface Γ_i (Figure 6.3). We formulate coupling conditions on each Γ_i. Moreover, we assume no-flow/no-slip condition for the free flow at the location of solid grains (no intersecting pore throat). This results in the following coupling conditions for the free-flow/pore-network model

$$\mathbf{v}_{\mathrm{ff}} \cdot \mathbf{n} = \mathbf{v}_{\mathrm{pm}} \cdot \mathbf{n} \qquad \text{on } \Gamma_i \ , \tag{6.13}$$

$$p_{\mathrm{pm}} = p_{\mathrm{ff}} \qquad \text{on } \Gamma_i \ , \tag{6.14}$$

$$\mathbf{v}_{\mathrm{ff}} \cdot \boldsymbol{\tau} = \begin{cases} v_{\mathrm{slip}} & \text{on } \Gamma_i \ , \\ 0 & \text{else} \ , \end{cases} \tag{6.15}$$

with

$$v_{\mathrm{slip}} = -\frac{1}{\beta_{\mathrm{pore}}} \left[(\nabla \mathbf{v} + \nabla \mathbf{v}^T) \mathbf{n} \cdot \boldsymbol{\tau} \right]_{\mathrm{ff}} + \left[\mathbf{v} \cdot \boldsymbol{\tau} \right]_{\mathrm{pm}} . \tag{6.16}$$

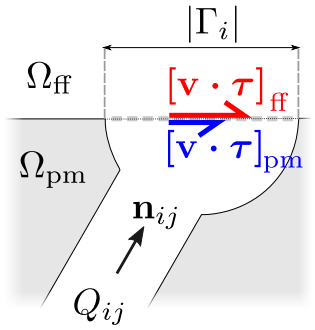


Figure 6.3: Schematic representation of a local interface for the free-flow/PNM.

We approximate the tangential component of the pore-body interface velocity as

$$[\mathbf{v} \cdot \boldsymbol{\tau}]_{\mathrm{pm}} = \frac{Q_{ij}}{|\Gamma_i|} [\mathbf{n}_{ij} \cdot \boldsymbol{\tau}]_{\mathrm{pm}} \ , \tag{6.17}$$

where $\mathbf{n}_{ij}$ is a unit normal vector parallel to the throat's central axis and pointing towards the interface Γ_i. The volume flow through the pore throat ij is given by Q_{ij} and $|\Gamma_i|$ is the area of the discrete coupling interface. Equations (6.15) and (6.16) can

be seen as the pore-scale analog to Equation (6.9) with a pore-local slip coefficient β_{pore}, which is determined numerically in a preprocessing step. See Weishaupt et al. [2020] for more details.

The three sets of coupling conditions (6.7)–(6.9), (6.10)–(6.12) and (6.13)–(6.15) are discretized corresponding to the adjacent subdomain models' discretizations, and the resulting coupled discrete models are treated by a monolithic strategy, assembling all contributions in a single system of equations for each model.

6.2.4 Benchmark scenarios

Corresponding to Figures 6.1 and 6.4, we investigated three scenarios with different flow problems induced by varying positions of inflow opening and solid inclusions. We considered laminar fluid flow through the coupled flow domain with viscosity $\mu = 10^{-3}$ Pa $\cdot$ s. In what follows, Section 6.2.4 describes the geometrical configuration and the boundary conditions, followed by a description of the system response quantities in Section 6.2.4 and the uncertainties in Section 6.2.4.

Geometrical setting and boundary conditions

We considered the free-flow region $\Omega_{\text{ff}} = (0, L) \times (\gamma, H)$ and the porous-medium domain $\Omega_{\text{pm}} = (0, L) \times (0, \gamma)$ with $L = 10.25\,\text{mm}$ and $H = 6\,\text{mm}$, separated by the sharp fluid–porous interface $\Gamma = (0, L) \times \{\gamma\}$, where the value for γ is uncertain.

For the first case (*case 1*), we investigated rectangular solid inclusions of size d and studied a flow problem where the flow is arbitrary to the fluid–porous interface Γ. The porous medium is isotropic, $\mathbf{K} = k\mathbf{I}$, and consists of 20×10 square solid inclusions of size $d = 0.25\,\text{mm}$ (Figure 6.4) leading to porosity $\phi = 0.75$. The inclusions are positioned in such a way that the line tangent to the top of the upper row of solid inclusions is given by $(0, L) \times \{5\,\text{mm}\}$ and the characteristic pore size appearing in coupling condition (6.12) is $\ell = 0.5\,\text{mm}$.

For the classical interface conditions (6.7)–(6.9) (*Classical IC*) the Beavers–Joseph parameter is typically taken $\alpha_{\text{BJ}} = 1$ in the literature, although it is often not the optimal choice [Rybak et al., 2021, Mierzwiczak et al., 2019, Lācis et al., 2020]. Here, we consider

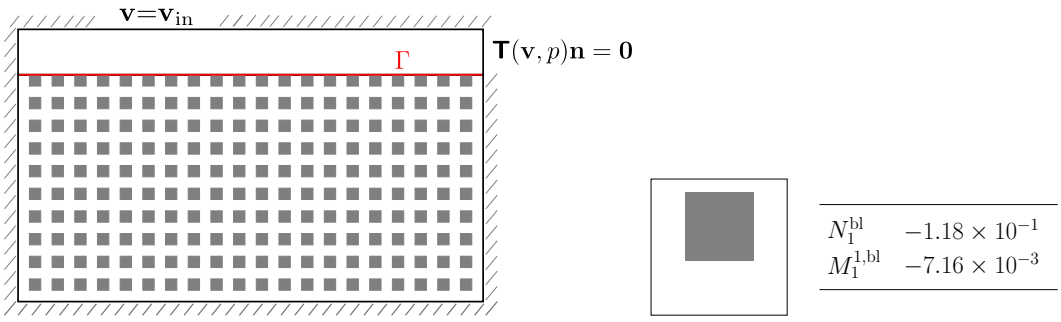


Figure 6.4: Schematic description of the coupled flow problem (left), unit cell and non-dimensional effective parameters for the interface location $\gamma = 5.05\,\mathrm{mm}$ (right).

α_{BJ} as an uncertain parameter, which is quantified in Section 6.2.4. The boundary layer constants appearing in the generalized coupling conditions (6.10)–(6.12) (*Generalized IC*) are computed numerically based on the geometrical configuration of the interfacial zone and are presented in Figure 6.4 (right). For details on the computation of these effective parameters, see Eggenweiler and Rybak [2021]. Note that the boundary layer constants N_1^{bl} and M_1^{bl} (Figure 6.4, right) are non-dimensional. For isotropic porous media, the constants $M_1^{2,\mathrm{bl}} = 0$ and $N_s^{\mathrm{bl}} = 0$, therefore, they do not appear in Figure 6.4 (right).

In order to obtain a closed formulation for the pore-scale problem (6.1)–(6.2), we set the following boundary conditions on the external boundary

$$\mathbf{v} = \mathbf{v}_{\mathrm{in}} = (0, V^{\mathrm{top}} \sin(\tfrac{1000}{3}\pi x)) \qquad \text{on } \Gamma_{\mathrm{in}}, \tag{6.18}$$

$$\mathbf{T}(\mathbf{v}, p)\, \mathbf{n}_{\mathrm{ff}} = \mathbf{0} \qquad \text{on } \Gamma_{\mathrm{out}}, \tag{6.19}$$

$$\mathbf{v} = \mathbf{0} \qquad \text{on } \Gamma_{\mathrm{nf}}, \tag{6.20}$$

where the inflow boundary $\Gamma_{\mathrm{in}} = (3\,\mathrm{mm}, 6\,\mathrm{mm}) \times \{H\}$, $\Gamma_{\mathrm{out}} = \{L\} \times (5.5\,\mathrm{mm}, H)$, $\Gamma_{\mathrm{nf}} = \partial\Omega \setminus (\Gamma_{\mathrm{in}} \cup \Gamma_{\mathrm{out}})$ and $\mathbf{n}_{\mathrm{ff}}$ is the outward unit normal vector on $\partial\Omega_{\mathrm{ff}}$. The boundary conditions for the coupled flow problem are presented schematically in Figure 6.4 (left).

The corresponding boundary conditions for the REV-scale model formulation given by (6.3)–(6.4) together with either the *Classical IC* (6.7)–(6.9) or the *Generalized*

IC (6.10)–(6.12) read

$$\mathbf{v}_{\mathrm{ff}} = (0, V^{\mathrm{top}} \sin(\tfrac{1000}{3}\pi x)) \qquad \text{on } \Gamma_{\mathrm{in}}, \tag{6.21}$$

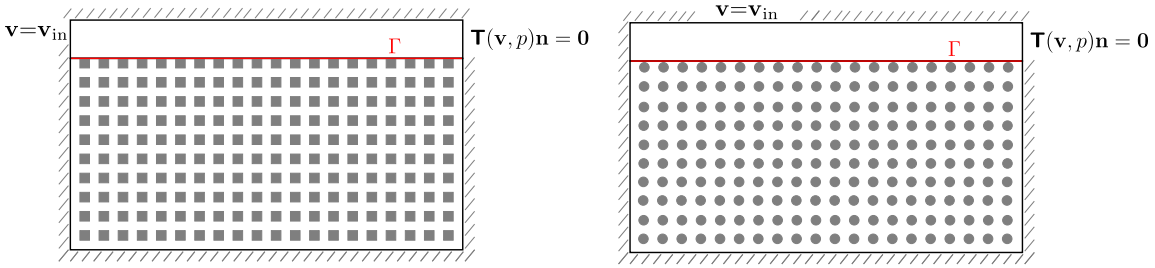
$$\mathbf{T}(\mathbf{v}_{\mathrm{ff}}, p_{\mathrm{ff}})\, \mathbf{n}_{\mathrm{ff}} = \mathbf{0} \qquad \text{on } \Gamma_{\mathrm{out}}, \tag{6.22}$$

$$\mathbf{v}_{\mathrm{ff}} = \mathbf{0} \qquad \text{on } \Gamma_{\mathrm{nf,ff}}, \tag{6.23}$$

$$\mathbf{v}_{\mathrm{pm}} \cdot \mathbf{n}_{\mathrm{pm}} = 0 \qquad \text{on } \Gamma_{\mathrm{nf,pm}}, \tag{6.24}$$

where $\Gamma_{\mathrm{nf,ff}} = \partial\Omega_{\mathrm{ff}} \setminus (\Gamma_{\mathrm{in}} \cup \Gamma_{\mathrm{out}} \cup \Gamma)$, $\Gamma_{\mathrm{nf,pm}} = \partial\Omega_{\mathrm{pm}} \setminus \Gamma$ and $\mathbf{n}_{\mathrm{pm}}$ denotes the unit normal vector on $\partial\Omega_{\mathrm{pm}}$ pointing outward the porous medium. The boundary conditions (6.21)–(6.23) also hold for the hybrid-dimensional free-flow/pore-network model (*Pore-Network*), such that no mass enters or leaves the domain through the pores on $\Gamma_{\mathrm{nf,pm}}$. The coupling conditions (6.13)–(6.15) are set on Γ for PNM.

We have also investigated two other cases: one with the opening boundary condition on the left side (*case II*) and another with circular inclusions in the porous medium (*case III*). These two cases both have similar boundary conditions to the *case I*. Figure 6.5 shows the configuration of the two cases.



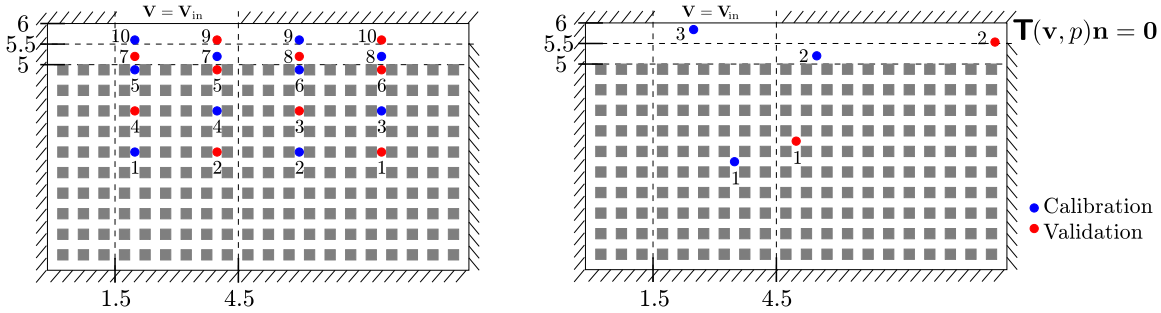
(a) *case II* (b) *case III*

Figure 6.5: Schematic configuration of case studies II and III.

System response quantities

A fundamental ingredient of each benchmark is the definition of so-called system response quantities (SRQs). These quantities define the prescribed output from the reference data as well as from the computational models that are compared via the validation metric. The SRQs can be either local or global quantities. While the former can take quantities within the solution domain on the PDEs, such as dependent

variables of the PDEs, the latter represents integral quantities or net flux out of a system. As part of model validation, we seek to compare system responses generated by different coupling concepts with the ones from the pore-scale resolved model (Section 6.2.1). Figure 6.6 shows the data extraction points for the velocity field (top) and the pressure (bottom). Since less variability is expected for pressure values, we have selected fewer extraction points for pressure responses. We train the surrogate models for all computational models based on the simulation results for the marked points. The points colored in blue and red provide the corresponding data for the calibration and validation steps, respectively. The positions of these points are identical for all investigated cases.



(a) Velocity field (b) Pressure field

Figure 6.6: Data extraction points for the calibration and validation scenarios.

The pore-scale resolved simulation results contain both macroscopic and microscopic details of the flow field. The latter becomes visible as oscillations of the pore-scale solutions in the porous medium. To make numerical simulation results comparable, we need to average them at the pore and REV scale. We consider volume averaging, where the averaged velocity field at a given point $\mathbf{x}_0 \in \Omega$ is obtained as

$$\mathbf{v}_{\text{avg}}(\mathbf{x}_0) = \frac{1}{|V(\mathbf{x}_0)|} \int_{V_{\text{f}}(\mathbf{x}_0)} \mathbf{v}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \,, \tag{6.25}$$

where $V(\mathbf{x}_0)$ is the representative elementary volume corresponding to $\mathbf{x}_0$ and $V_{\text{f}}(\mathbf{x}_0)$ is its fluid part. The representative elementary volume $V(\mathbf{x}_0)$ has the same size as the periodicity cell Y^{ℓ}. Moreover, the simulation results in the free-flow region need to be averaged correspondingly such that the interpretation of the SRQs is the same in all cases.

Uncertainties and errors

So far, various coupling concepts for free-flow and porous-medium flow have been presented. The uncertainty, due to the choice of adequate representation of the system of interest, is known as conceptual uncertainty. In addition to the conceptual uncertainty, each computational model has uncertain parameters, such as material parameters or interface location, requiring a thorough investigation. This type of uncertainty is known as parametric uncertainty. Uncertain model inputs, defined later, must be propagated through the model or simulation (also known as uncertainty propagation) to effectively assess response quantities of competing modeling concepts and validate the corresponding computational models against a reference solution.

The Stokes–Darcy problem (6.3)–(6.4) with the *Classical IC* (6.7)–(6.9) contains four uncertain model parameters: the maximum boundary velocity at the inflow boundary V^{top} (Figure 6.4), the exact interface position γ, the permeability tensor $\mathbf{K} = k\mathbf{I}$ and the Beavers–Joseph slip coefficient α_{BJ}. The exact location of the fluid–porous interface for REV-scale models is not known *a priori*. The Beavers–Joseph coefficient α_{BJ} is supposed to contain the information on the surface roughness [Beavers and Joseph, 1967, Le Bars and Worster, 2006]. An investigation to calibrate this parameter was recently carried out in Rybak et al. [2021], however, only for isotropic media. There was also an attempt to determine the Beavers–Joseph coefficient experimentally for flows parallel to the fluid–porous interface, isotropic and orthotropic porous media [Mierzwiczak et al., 2019], where the Beavers–Joseph parameter was found to be $\alpha_{\text{BJ}} < 1$ and dependent on the intrinsic permeability. Finally, the permeability tensor appearing in the Beavers–Joseph condition (6.9) is not necessarily the permeability of the porous bulk, as in the standard models [Discacciati et al., 2002, Discacciati and Quarteroni, 2009], but could also be permeability of the near-interfacial region [Lācis and Bagheri, 2017, Zampogna and Bottaro, 2016].

In contrast to the *Classical IC*, the *Generalized IC* do not contain the Beavers–Joseph coefficient. Further, the *Generalized IC* rely on the assumption that the interface location may not be below the top of the solid inclusions. Correspondingly, the parameters and their associated distributions as prior knowledge for the Stokes–Darcy model with the *Classical IC* and the *Generalized IC* are listed in Tables 6.1 and 6.2, respectively.

As for the *Pore-Network* model, we considered the total conductance g_{ij} in (6.6) (see

Table 6.1: List of uncertain parameters and their defined distributions for the classical coupled Stokes–Darcy model.

Parameter name	Range	Unit	Distribution type
Boundary velocity, V^{top}	$[5 \cdot 10^{-4}, 1.5 \cdot 10^{-3}]$	m/s	uniform
Exact interface location, γ	$[4.9, 5.1]$	mm	uniform
Permeability, k	$[10^{-10}, 10^{-8}]$	m^2	uniform
Beavers–Joseph parameter, α_{BJ}	$[0.1, 4]$	-	uniform

Table 6.2: List of uncertain parameters and their associated distributions for the Stokes–Darcy model with the generalized interface conditions.

Parameter name	Range	Unit	Distribution type
Boundary velocity, V^{top}	$[5 \cdot 10^{-4}, 1.5 \cdot 10^{-3}]$	m/s	uniform
Exact interface location, γ	$[5.0, 5.1]$	mm	uniform
Permeability, k	$[10^{-10}, 10^{-8}]$	m^2	uniform

Figure 6.2) as uncertain parameter to be inferred during the calibration phase. This parameter plays the role of permeability in the pore-network setting. Another uncertain input parameter is the pore-scale slip coefficient β_{pore}. It can be determined numerically in a preprocessing step, in which it is approximated by solving a simplified, equivalent problem of free flow over a single pore throat intersecting with the lower boundary of the free-flow channel [Weishaupt et al., 2020]. The list of considered uncertain parameters and their associated distribution as prior knowledge for the PNM are presented in Table 6.3.

Table 6.3: List of uncertain parameters and their specifications for the pore-network model.

Parameter name	Range	Unit	Distribution type
Boundary velocity, V^{top}	$[5 \cdot 10^{-4}, 1.5 \cdot 10^{-3}]$	m/s	uniform
Total conductance, g_{ij}	$[10^{-7}, 10^{-5}]$	$\text{m}^3/(\text{s} \cdot \text{Pa})$	uniform
Pore-local slip coefficient, β_{pore}	$[10^3, 10^5]$	1/m	uniform

As opposed to uncertainties, errors are defined as the difference between the true value and the predicted value and have both a sign and a magnitude. We consider the errors associated with the model discrepancy error, numerical approximation, and surrogate modeling in our analysis. The analyzed models in our benchmark study could never perfectly reproduce the reference solution. This difference can be attributed to the

presence of model discrepancy. This discrepancy can be attributed to simplified assumptions, missing physics, upscaling due to scale differences. We parameterize it as $\Sigma(\theta_\epsilon)$ and treat its parameter θ_ϵ as an additional unknown parameter. Following Wagner et al. [2021], we infer these parameters jointly with model parameters θ_k in (2.6). We consider a diagonal covariance matrix as $\Sigma = \sigma^2 \mathbb{I}_{N_{\text{out}}}$ with a scalar unknown parameter σ^2 for each system response quantity, i.e., velocity and pressure (Section 6.2.4). N_{out} stands for the number of data points.

The governing equations of the models under investigation in this study require the approximation of numerical solutions. These approximations provide an additional source of error. Since quantifying numerical errors is the main focus in the verification of numerical schemes, we only investigate the discretization error that originates from a certain choice of the mesh size. Similar to previous chapter, we take a heuristic approach from Oberkampf and Roy [2010] to quantify this error, in that we fit generalized Richardson extrapolation (Equation (5.1)) to estimate the error by comparing three different mesh spacings.

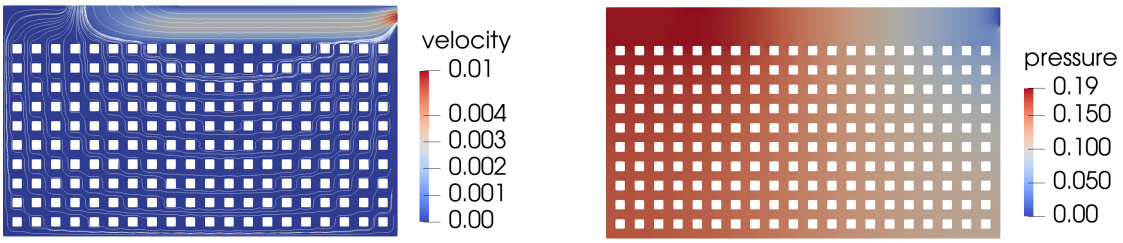
As previously mentioned, the computational models are substituted with the easy-to-evaluate surrogate models in the Bayesian analysis to offset the computational cost. This replacement also introduces a new source of error, known as a surrogate prediction error. Ignoring this error could result in a biased posterior distribution. As for prediction uncertainty, a mean squared error based on a testing set can provide a good estimate of the surrogate error variance. We incorporated the errors discussed above in the MCMC simulation method used in the calibration stage to approximate the posterior distribution. We directly sum up all the covariance matrices of errors to obtain the likelihood calculations' total covariance matrix Σ in the likelihood function in Equation (2.6). Here, we assume that all these errors follow a normal distribution and are independent of each other.

6.2.5 Solution procedure

In this study, we compared the coupled models (using either REV-scale formulation or *Pore-Network* model in Ω_{pm}) with the pore-scale resolved model (6.1)–(6.2). For the REV-scale model formulation, we consider the Stokes–Darcy problem with the *Classical IC* (6.7)–(6.9) and the *Generalized IC* (6.10)–(6.12). The pore-scale and

coupled models have been implemented in the open-source simulator DuMu^{x} [Koch et al., 2021a].

Here, our goal is to assess the coupled model's accuracy compared to the above-mentioned REV-scale approaches and under the influence of pore-scale parameter uncertainty. As reference data, we used the fully resolved pore-scale model for the velocity and pressure (Figure 6.7). However, it is worth mentioning that the Stokes–Darcy model with *Classical IC* and *Generalized IC* can only offer predictions on the REV scale. Therefore, we average the values of SRQs obtained for the fully resolved pore-scale model as well as the *Pore-Network* model for consistency. The averaging is performed via a volume averaging approach, discussed in Section 6.2.4 to make the REV-scale numerical simulation results comparable with that of the pore-scale resolved simulation.



(a) Velocity field (b) Pressure field

Figure 6.7: Streamlines of the pore-scale (reference) simulation.

Replacing the models with their surrogates drastically reduces the total computational time of the analysis. This gain is essential in computationally demanding uncertainty quantification tasks, such as propagation or inference. In this study, we observed that using a well-trained surrogate model could speed up one simulation run from $10 \sim 15$ s to only $0.005 \sim 0.007$ s with acceptable accuracy.

The predictive ability of the models has been analyzed via the surrogate-based Bayesian procedure. In the calibration phase, we updated the prior knowledge on the uncertain model parameters according to Section 6.2.4. We conditioned the responses of all analyzed models on the velocity and pressure values extracted from the pore-scale simulations that are marked as blue points in Figure 6.6. To do so, we employed MCMC approach to perform Bayesian inference Section 2.1.3 using surrogate representation,

presented in Chapter 3. We use an Affine Invariant Ensemble Sampler (AIES) to approximate the posterior distribution (Section 2.1.3).

We trained each surrogate model with the simulation outcomes of 300 numerical model runs to accelerate this Bayesian updating step. The AIES-MCMC sampler was run for an ensemble of 50 Markov chains on each surrogate. We monitor the convergence of the sampler using the integrated auto-correlation time, which estimates the number of evaluations of the posterior probability density function to draw independent samples from the target density [Sokal, 1997]. The MCMC sampler is run until the convergence criterion of 1% for the difference in the auto-correlation time between two consequent monitoring steps is met. We retrain a new set of surrogate models in the validation stage based on the updated parameter distribution (posterior distribution) obtained after calibration. With these surrogate models, we propagate the posterior parametric uncertainty to estimate the posterior predictive distribution of models to be passed to the Bayesian metric calculation step. The abovementioned procedures have been performed using the Python package *BayesValidRox* (Chapter 4).

As discussed in Section 6.2.4, using surrogates may introduce additional errors to the inference process. To include this error, we test the surrogate models with 150 simulation runs (test sets) which are different from the training sets. Comparing the surrogates' prediction with the results from the test sets, we observed a considerably low validation error between 10^{-8} and 10^{-11} for all models, indicating an acceptable prediction accuracy. Moreover, we estimated Mean Square Error (MSE) for each surrogate model, which is a good estimate of the surrogate error variance [Xu and Valocchi, 2015]. When evaluating the likelihood $p(\mathcal{Y}|M_k, \theta_k)$ in (2.6), we add a diagonal matrix Σ_{PCE} with elements $\sigma^2_{\text{PCE},i} = MSE_i$, $i = 1, 2, ..., N_{\text{out}}$ to Σ, assuming that the surrogate errors are independent and follow a normal distribution with zero mean. Moreover, following Schöniger et al. [2015b], we perturbed the reference data with some additive noise to account for uncertainty associated with the BME values, the resulting Bayes factors, and posterior model weights. With this approach, we investigated the impact of other possible sources of errors on the validation metrics that are not considered in the calculations.

As mentioned in Section 6.2.4, we jointly inferred the uncertain parameters with the scalar unknown parameters σ^2_{vel} and σ^2_p for each system response quantity, i.e., velocity

and pressure. For these parameters, we assumed uniform distributions $\sigma^2_{vel} \sim \mathcal{U}[0, 10^{-5}]$ and $\sigma^2_p \sim \mathcal{U}[0, 10^{-3}]$ as priors, for velocity and pressure, respectively.

6.3 Results and discussion

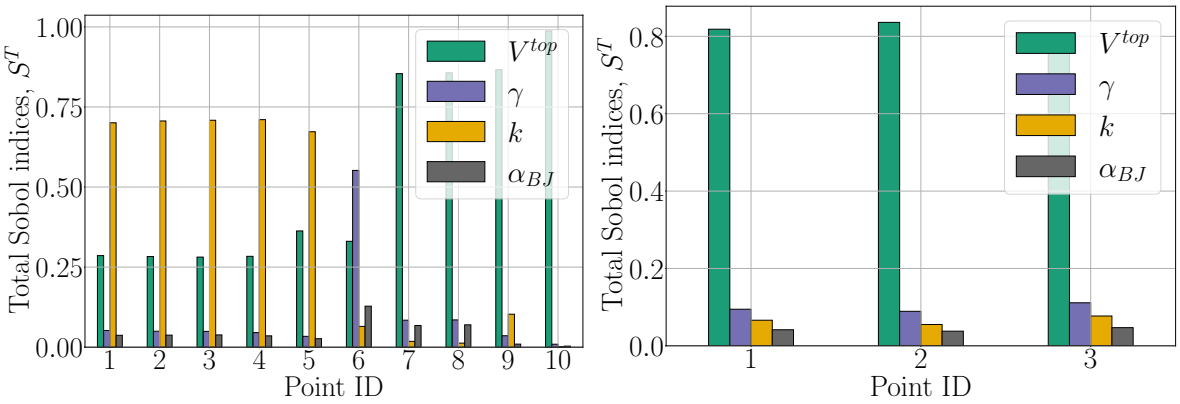
The current section offers insights into the analysis of predictive abilities in Section 6.3.2 and model comparison in Section 6.3.3 using the surrogate-based Bayesian validation framework. Additionally, we assessed the influence of various modeling parameters on the final model prediction, performing the global sensitivity analysis in Section 6.3.1.

6.3.1 Global sensitivity analysis

In this section, we analyze how the variability of the model response quantities introduced in Section 6.2.4 at the selected data extraction points (Figure 6.6) is affected by the variability of each input variable or combinations thereof. This is achieved via sensitivity analysis. Section 3.1.7 explored the connection of polynomial representation to global sensitivity measures [Oladyshkin et al., 2012] and used the so-called *Sobol indices* [Sobol', 1993], derived from a variance decomposition of model outputs in terms of contributions of each input parameter or combinations thereof. In what follows, we present the total Sobol indices for the SQRs and the data extraction points defined in Section 6.2.4, for all three models for the calibration scenario.

Classical IC In Figure 6.8, we provide the total Sobol indices for the calibration points in blue (Figure 6.6) for velocity and pressure for the *Classical IC* (6.7)–(6.9). We observe that the boundary velocity V^{top} has the most contribution to the velocity variance (Figure 6.8, left) in the free-flow region and pressure field (Figure 6.8, right) for the analyzed points in the domain. Moreover, the exact interface location plays an important role for the velocity field, especially near the interface (Figure 6.8, left) and influences the pressure field as well (Figure 6.8, right). The value of the Beavers–Joseph parameter has a higher impact on the velocity near the interface than other parts of the domain. However, this parameter does not play an essential role for the pressure. The

permeability k significantly affects the velocity in the porous-medium domain, whereas its influence on the velocity in the free-flow region and the pressure field is small.

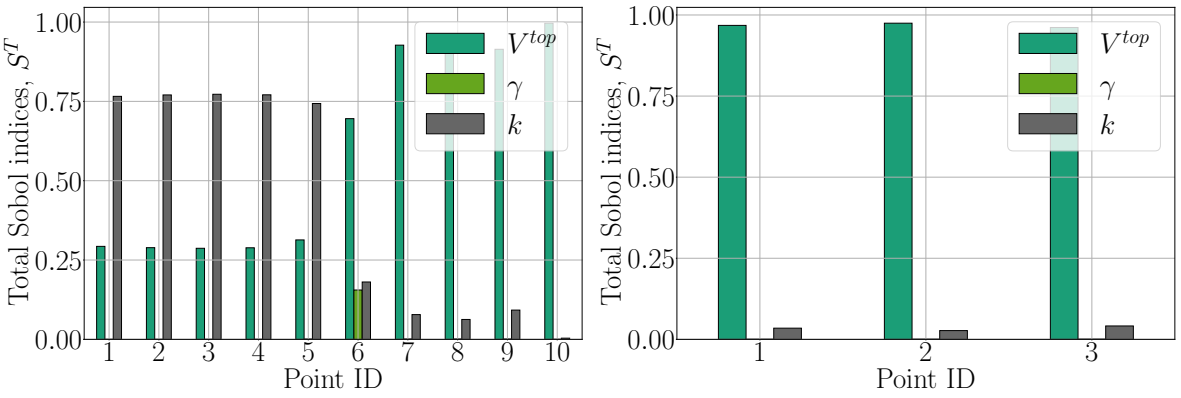


(a) Velocity (b) Pressure

Figure 6.8: Total Sobol indices of the Stokes–Darcy model with the *Classical IC* for the calibration (blue) points in Figure 6.6.

Generalized IC In Figure 6.9, the total Sobol indices for velocity and pressure before calibration (based on prior distributions of parameters) are presented for the selected blue points in Figure 6.6. For the *Generalized IC*, the information about the exact interface location γ is included in the boundary layer constants N_1^{bl} and $M_1^{1,\text{bl}}$ appearing in condition (6.12). Therefore, the exact position of the interface does not influence the overall system behavior in comparison to the *Classical IC*. The permeability k (in the porous-medium) and the inflow velocity V^{top} have a significant impact both on the velocity (in the free-flow region) and the pressure field, as in the case of the *Classical IC*.

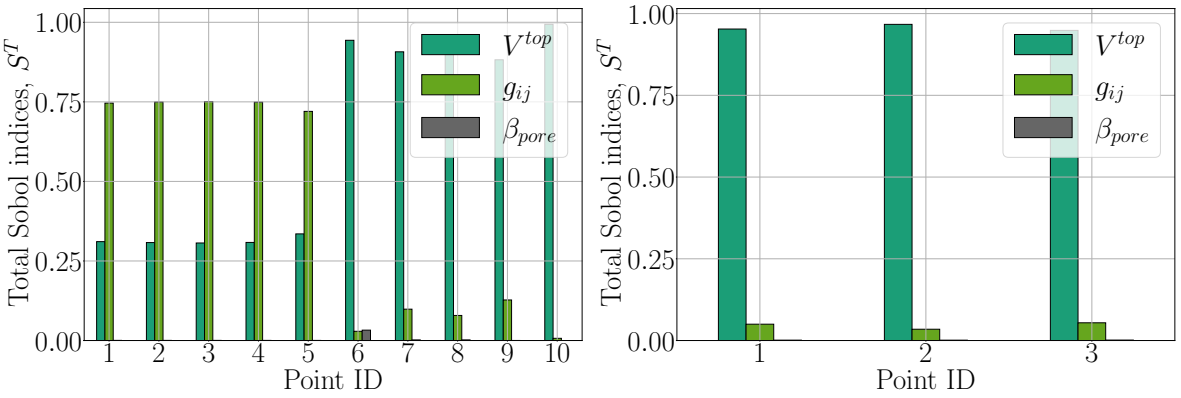
Pore-network model Figure 6.10 shows the total Sobol indices for velocity (left) and pressure (right) for the blue points in Figure 6.6. As for the REV-scale coupled models, we observe a dominant influence of V^{top} for all points. As expected, the influence of the total conductance is more prominent in the porous domain, which is comparable to the influence of permeability for the REV-scale coupled models. The influence of the pore-scale slip parameter β_{pore} shows a relatively small influence on the variability of velocity at point 6 and is hardly visible at other locations in the free-flow region. This matter is most likely because the slip coefficient only affects the flow field in the free-flow domain



(a) Velocity (b) Pressure

Figure 6.9: Total Sobol indices of the Stokes–Darcy model with the *Generalized IC* for the calibration points.

Ω_{ff} very locally, directly above the interface pore. However, the averaging volume used for the evaluation takes into account a larger portion of the free-flow region, where the influence of β_{pore} is fairly small. In analogy to the REV models, V^{top} also has a dominating influence on the pressure and velocity in the free-flow region.



(a) Velocity (b) Pressure

Figure 6.10: Total Sobol indices of the *Pore-Network* model for the calibration points.

6.3.2 Analysis of predictive abilities

In this section, we present the result of the analysis of the predictive ability of all three discussed conceptual models by showing their parametric posterior and the corresponding predictive distributions. In what follows, we present the parameters' updated

(posterior) distribution and model discrepancy errors after calibration obtained by the MCMC sampler for all three models. Afterward, a figure containing the posterior predictive of models next to each other versus the reference data is provided.

Classical IC Figure 6.11 presents the posterior distribution obtained via the Bayesian inference using the calibration (blue) points in Figure 6.6. The 50 percent quantiles, alongside the 15 and 85 percent quantiles, are displayed on top of the histograms shown in the diagonal plots. Most posterior distributions of the parameters follow a Gaussian distribution. However, the distribution of the interface location γ and the Beavers–Joseph slip coefficient α_{BJ} exhibits a long right tail. Moreover, a slight correlation between γ and α_{BJ} is observed.

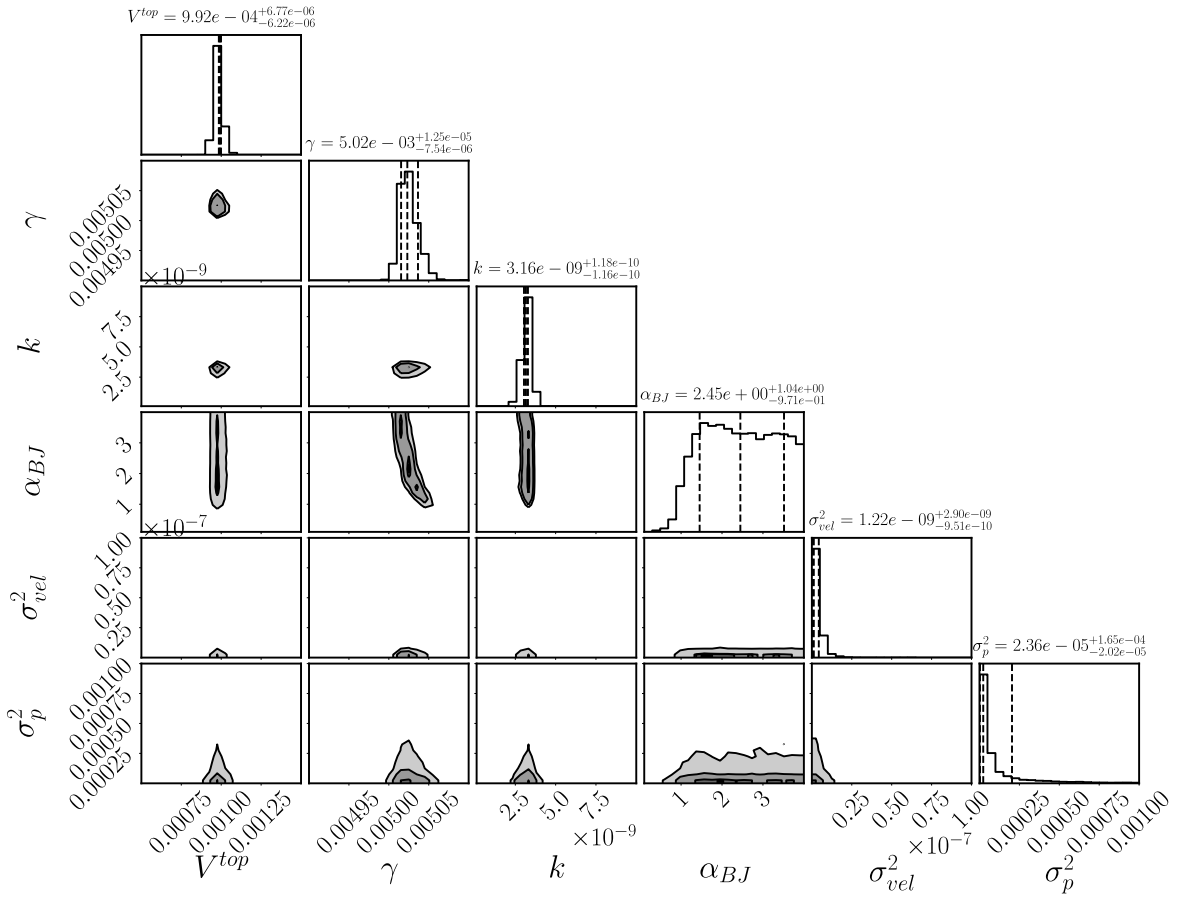


Figure 6.11: Posterior parameter distribution of the Stokes–Darcy model with the *Classical IC* after calibration to the reference data from the pore-scale model.

Generalized IC Similar to the procedure described above, the surrogate-based Bayesian calibration offers insight into the posterior distributions of modeling parameters for the Stokes–Darcy model with the *Generalized IC* (Figure 6.12). As opposed to the *Classical IC*, the interface location γ for this coupling condition shows a slightly wider distribution. This observation indicates that the exact position of interface does not influence the overall system behavior in comparison to the *Classical IC*.

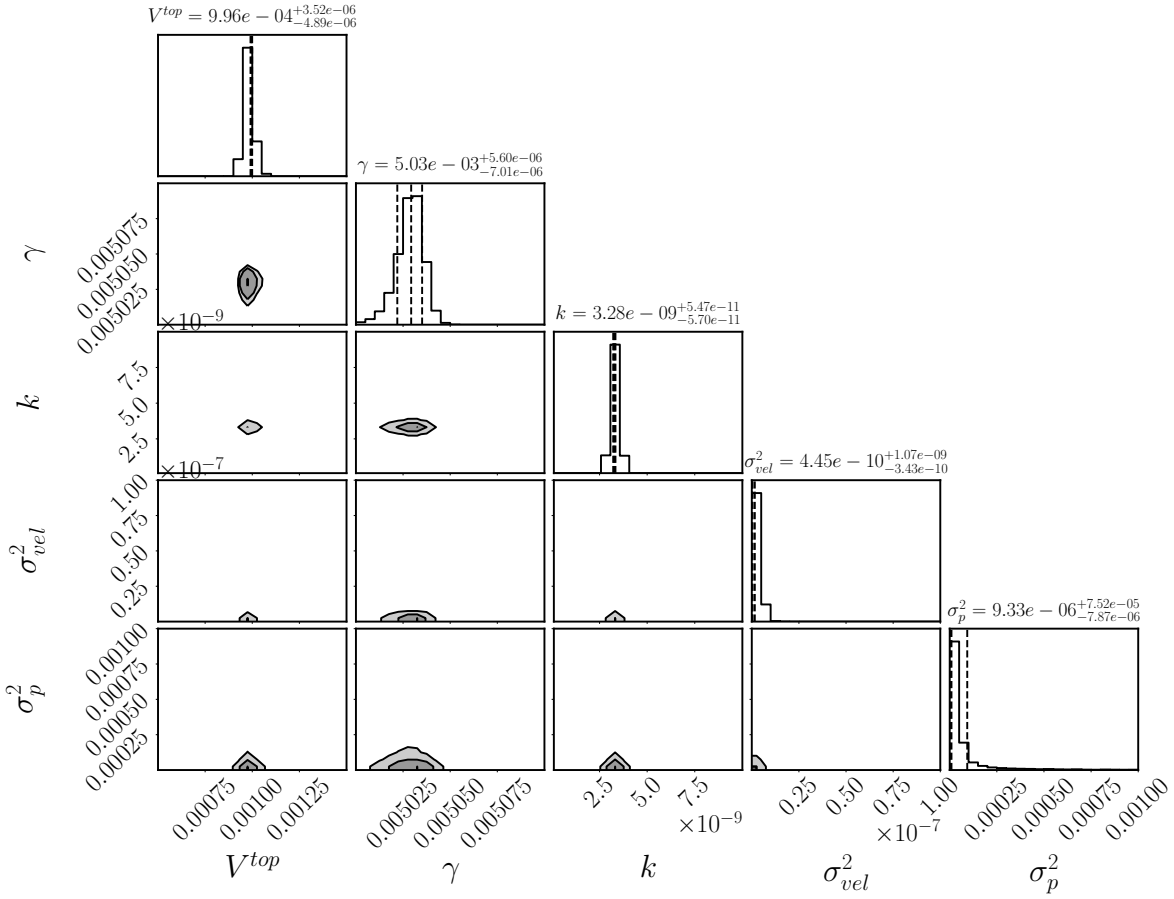


Figure 6.12: Posterior parameter distribution of the Stokes–Darcy model with the *Generalized IC* after calibration to the reference data from the pore-scale model.

Pore-network model For the *Pore-Network* model, we also have used the calibration (blue) points (Figure 6.6) to perform surrogate-based Bayesian inference. Figure 6.13 illustrates the posterior parameter distribution of the *Pore-Network* model. The β_{pore} distribution covers a wider range. This issue can be attributed to the insensitivity of the model results to this parameter, as presented by the total Sobol indices in Figure 6.10.

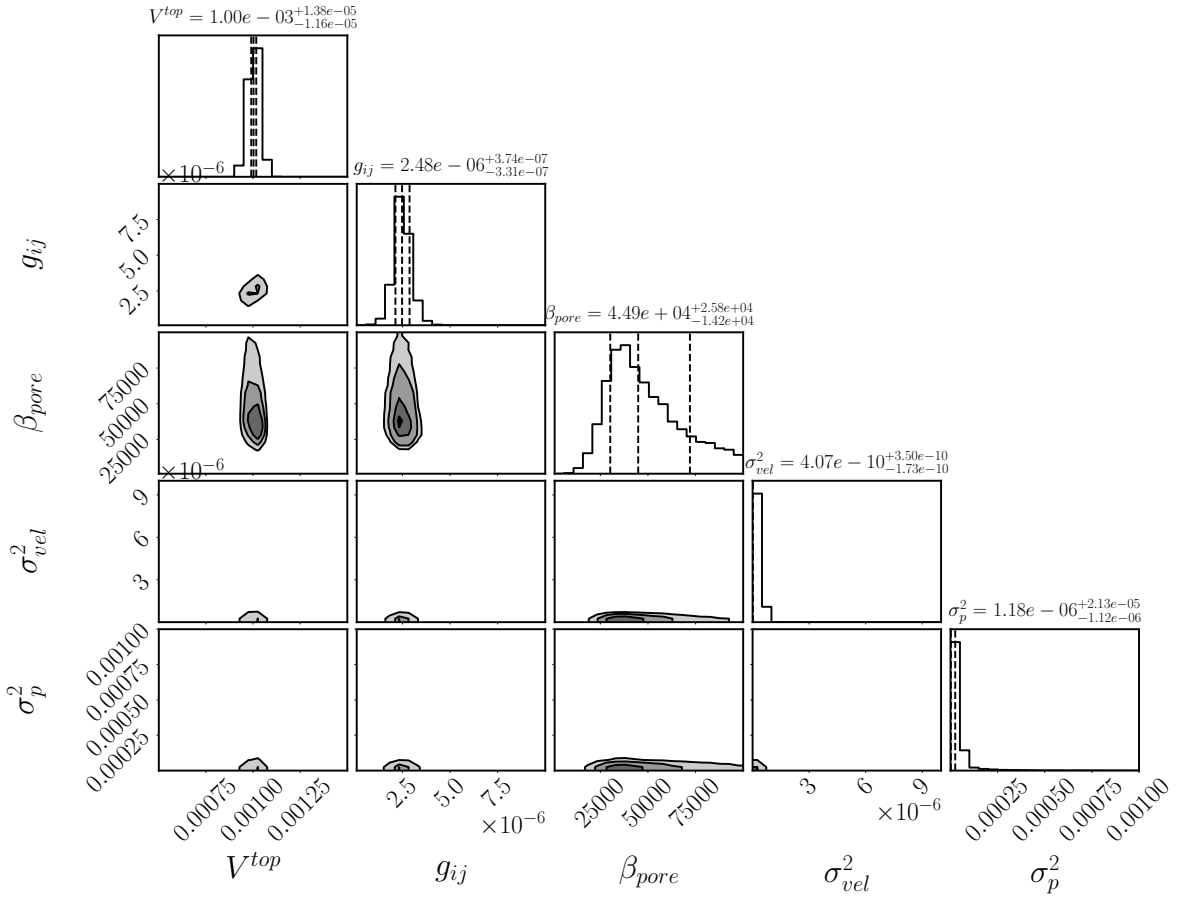


Figure 6.13: Posterior parameter distribution of the *Pore-Network* model after calibration to the reference data from the pore-scale model.

Posterior predictives To obtain the models' posterior predictive distributions, we need to propagate the posterior parametric uncertainty presented so far through the models. The result offers a possibility of analyzing how post-calibration uncertainty affects the SRQs. To perform the post-calibration uncertainty propagation, we have trained a new surrogate for each competing model using new training sample points drawn from the posterior parameter distribution. For better visual comparison, we plot the posterior predictive of models next to each other. Figures 6.14 and 6.15 illustrate the mean and standard deviations of the model predictive distributions in a bar chart for the velocity and pressure response quantities, respectively.

In particular, Figure 6.14 reveals that all analyzed models provide accurate predictions at the points located in the deeper part of the porous medium (1 to 4). However, the predictions at the points near the interface (5 to 8) suggest that the Stokes–Darcy model

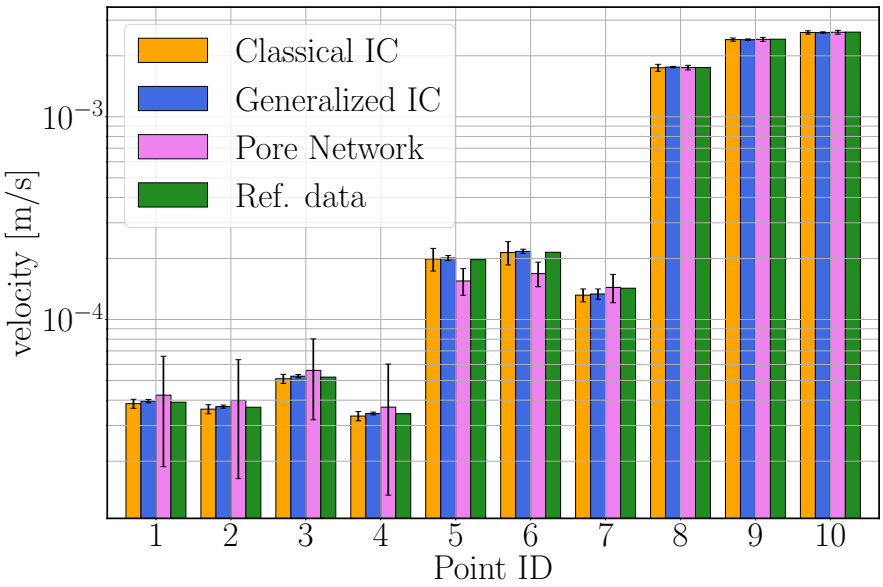


Figure 6.14: The velocity predictions of all models in the validation step against the reference data from the pore-scale model.

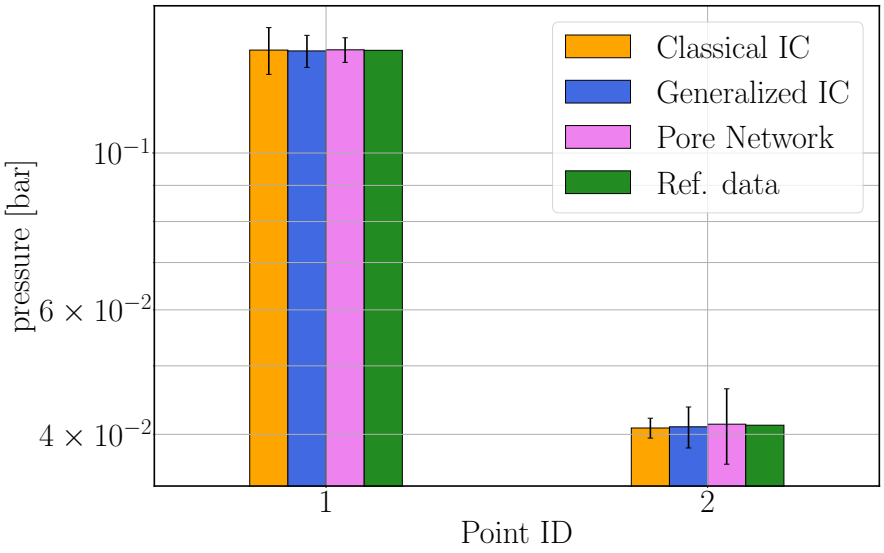


Figure 6.15: The pressure predictions of all models in the validation step against the reference data from the pore-scale model.

with *Classical IC* and *Generalized IC* provide more accurate predictions than the *Pore-Network* model. The REV-scale model with *Generalized IC* shows less uncertainty, i.e., lower standard deviation, in its prediction at the vicinity of the interface between the porous medium and the free-flow. Moreover, Figure 6.15 confirms that all models are able to provide accurate pressure values.

6.3.3 Model comparison

We perform the model comparison using the so-called posterior model weights via the BMS explained in Section 2.2.2. Such an analysis offers an aggregated comparison of a model's outputs to the validation set of reference data from the pore-scale model that are marked in red in Figure 6.6. We used the newly constructed surrogate representation during the validation stage for model comparison analysis to compute the BME values in Equation (2.8). These values are required to calculate the posterior model weights in Equation (2.23) and the Bayes factors in Equation (2.21). Additionally, the use of the advanced surrogate representation provides a possibility to assess the uncertainty of the BME values and the corresponding model weights. Table 6.4 presents a detailed statistical summary of the model weights and provides a ranking. It also reports the information regarding the post-calibration uncertainty with the help of the deviation regarding 25% and 75% percentiles.

Table 6.4: Statistical summary of posterior model weights after validation.

Model	Model weights	Rank
Classical IC	$0.003^{+0.002}_{-0.001}$	2
Generalized IC	$0.997^{+0.001}_{-0.002}$	1
Pore-network	$0.000^{+0.000}_{-0.000}$	3

The expected model weights under noisy pore-scale data assumption convey a relatively clear model ranking in favor of *Generalized IC*, with *Classical IC* as second and the *Pore-Network* model ranking last. It is worth mentioning that the model weights close to zero for the *Classical IC* and *Pore-Network* models can be attributed to the high prediction uncertainty of these models. This fact is represented by the error bars in Figure 6.14. Moreover, a considerable mismatch can be detected between the expected velocity prediction of the *Pore-Network* model and the reference data at validation points 5 and 6. As for the *Classical IC*, the velocity prediction uncertainty is higher than that of the *Generalized IC*. This difference is mainly for the points in the interface's vicinity and the free-flow region.

Assessments of confidence in the model ranking have been investigated employing BHT for pairwise comparison of models based on the validation scenario. In the introduced uncertainty-aware Bayesian validation framework, Bayes Factors provide an objective

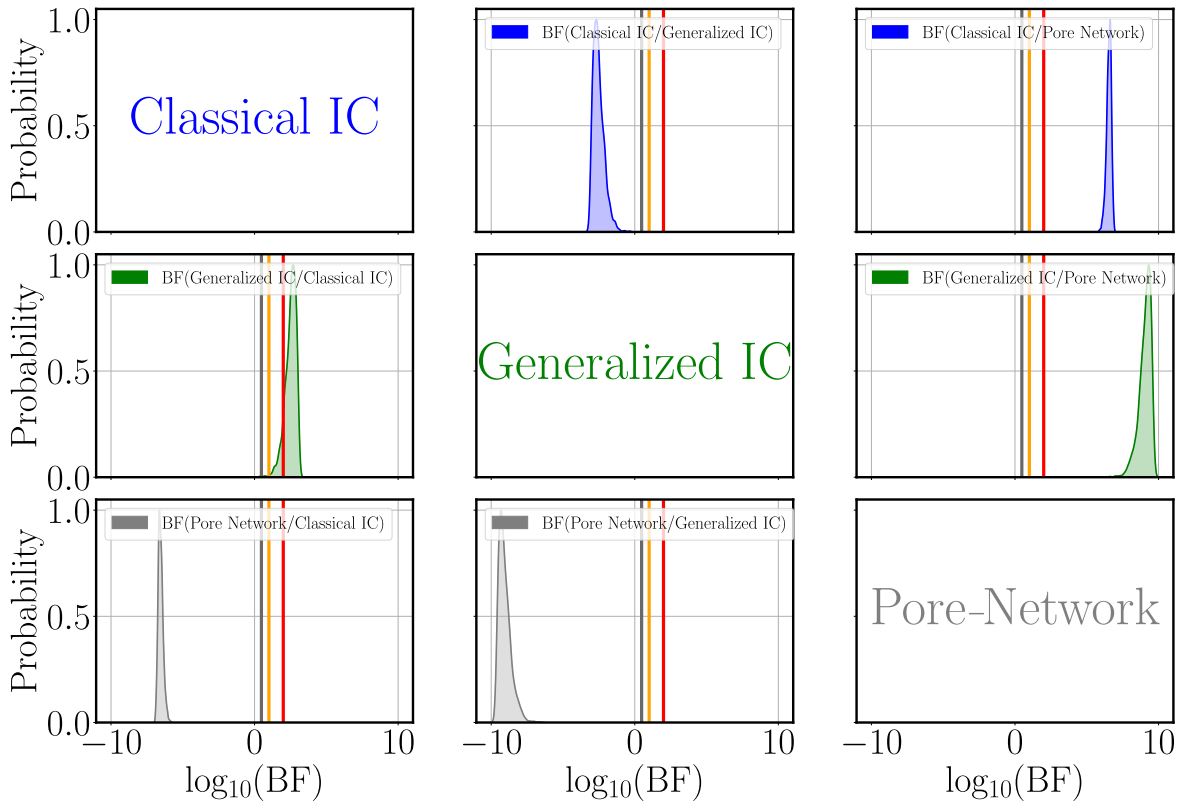


Figure 6.16: Distributions of $\log_{10}$ (Bayes Factor) for the pairwise comparison of competing models based on the validation scenario.

measure of significance that quantifies the evidence in favor of one model's superiority against another. Figure 6.16 presents the probability density functions of $\log_{10}(\text{BF})$ all perturbed velocity and pressure data sets in a three-by-three matrix. Here, we compute three Bayes Factors for each model against its counterpart. The significant levels in a $\log_{10}$-scale, introduced in Jeffreys [1961] are marked with vertical lines. Gray lines represent equally strong evidence for both models. Orange and red lines indicate thresholds for strong and decisive evidence in favor of one model against the other, respectively.

The first plot in the second row in Figure 6.16, e.g., shows the distribution of $\log_{10}(\text{BF})$ in favor of *Generalized IC* against *Classical IC*. This plot reveals that for most of the perturbed data sets, the Bayes factor is in the region where decisive evidence ($\log_{10}(\text{BF})$ greater than two) exists in favor of *Generalized IC* to outperform *Classical IC*. Similarly, in all the analyzed cases (perturbed data sets), *Classical IC* could be clearly favored against *Pore-Network* model based on the decisive evidence (the plot in the first row,

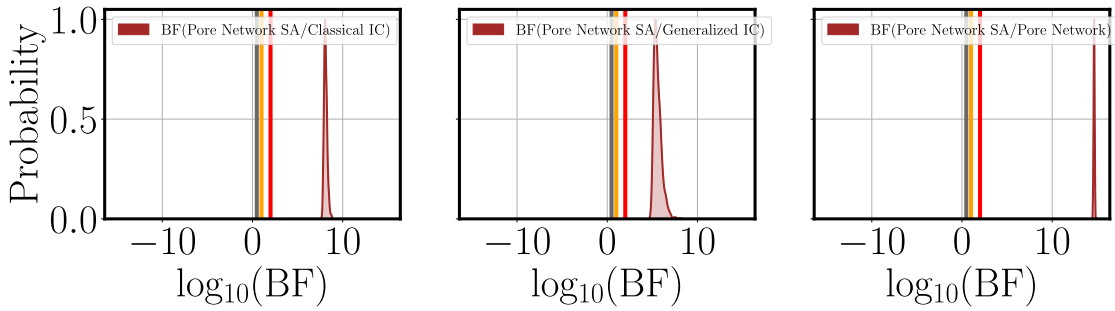


Figure 6.17: Distributions of $\log_{10}$ (Bayes factor) of the *Pore-Network* model, with the surface averaging against competing models based on the validation scenario.

the last column). Moreover, the distribution in the second row, the third column of Figure 6.16 reveals that the Bayes factor distribution of *Generalized IC* against *Pore-Network* model proves decisive evidence in favor of *Generalized IC*.

The results presented so far are based on a comparison of the SRQs with averaged SRQs of the fully resolved Stokes simulation, as the Stokes–Darcy model with *Classical IC* and *Generalized IC* could offer a prediction on the REV scale only. However, one could directly compare the *Pore-Network* model to the reference data at the pore scale without performing volume averaging by calculating the surface-averaged pore-scale velocity at the pore-throat cross-sections. We denote the *Pore-Network* model with the pore-throat surface averaging model as the *Pore-Network SA* model and its Bayes factors distribution is shown in Figure 6.17. The velocities of *Pore-Network SA* model are not defined within the pore bodies but only at the pore throats, which explains why the results of Figure 6.17 show stronger evidence in favor of the *Pore-Network SA* model compared to the other concepts. Therefore, the *Pore-Network SA* model avoiding additional averaging steps is a suitable approach when detailed pore-scale information is considered. Alternatively, the Stokes–Darcy model with *Generalized IC* adequately represents the underlying physical processes once only the REV-scale information is available.

In Section 2.2.2, we introduced an upper limit for the BME value via TOM. In what follows, I compare the models' BME distributions with the optimal distribution of BME (obtained for the TOM) by quantitative measures. Since likelihoods tend to show the largest mass around zero, the BME distributions used in this quantitative comparison are transformed in $\log_{10}$ scale, as it is more intuitive. To obtain the probability density

functions (PDFs) of log-BME, $p(Y)$ with $Y = \log_{10}(\text{BME})$, can be obtained from the ensemble of 10,000 replications of BME per model by kernel density estimation [Bowman and Azzalini, 1997]. The difference between PDFs of log-BME can be estimated with various distance measures. The following two measures are used for this study:

- Mode distance (D_{mode}): It measures the distance between the modes $\tilde{Y} = \max_\gamma p(Y)$ of the respective log-BMD densities:

$$\mathrm{D}_{mode}(\mathcal{M}_k, \mathcal{M}_l) = \tilde{Y}_{\mathcal{M}_k} - \tilde{Y}_{\mathcal{M}_l} \tag{6.26}$$

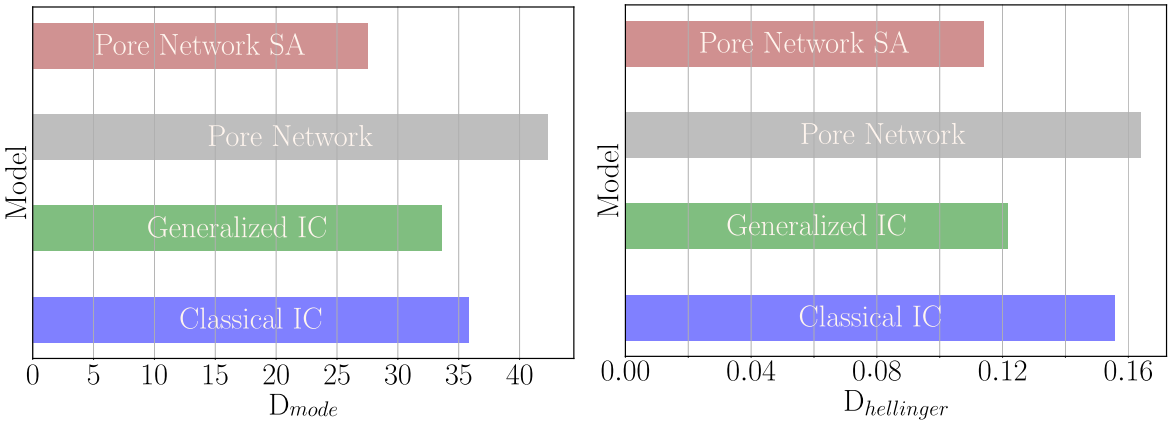
- Hellinger distance ($\mathrm{D}_{Hellinger}$): It is a dimensionless metric with fixed bounds of 0 (identical distributions) and 1 (no overlap at all) and defined as

$$\mathrm{D}_{hellinger}(\mathcal{M}_k, \mathcal{M}_l) = \sqrt{1 - BC\left(p(Y \mid \mathcal{M}_k), p(Y \mid \mathcal{M}_l)\right)}, \tag{6.27}$$

  where BC represents the Bhattacharyya coefficient, which reads:

$$BC\left(p(Y \mid \mathcal{M}_k), p(Y \mid \mathcal{M}_l)\right) = \int \sqrt{p(Y \mid \mathcal{M}_k) p(Y \mid \mathcal{M}_l)}. \tag{6.28}$$

Figure 6.18 shows the distance measures for the averaged case. Note that the TOM does not appear in the figure as its distance to itself would result in zero.



(a) D_{mode} (b) $\mathrm{D}_{hellinger}$

Figure 6.18: Comparison of $\log_{10}(\text{BME})$ distributions of models with that of TOM.

The distance measures in Figure 6.18 can also be interpreted as the remaining distance

or room for model improvement between the individual models and TOM. Both measures indicate that for the averaged case, the *Generalized IC* has the lowest distance and the *Pore Network* model has the highest distance to TOM. However, the *Pore-Network SA* model is the closest to the TOM when compared to the detailed pore-scale model.

In addition to the setup presented in Section 6.2.4, we also analyzed two other cases, shown in Figure 6.5. Firstly, we considered a setup with the same geometrical configuration, but the inlet boundary was located at the left domain edge in the free-flow region with an opening of 1.5 mm from the top (*case II*). This setup induces a flow profile parallel to the interface. Comparing *Classical IC* with *Generalized IC*, we witnessed no substantial evidence in favor of any model. This observation is in line with the results from Eggenweiler and Rybak [2021], where the authors showed that the Stokes–Darcy problem with *Classical IC* and *Generalized IC* provides similar simulation results for parallel flows to the porous layer. The second additional setup (*case III*) is based on the same flow models and boundary conditions as presented in Section 6.2.4, however, the solid inclusions are circular. We compared the Stokes–Darcy model with *Classical IC* and *Generalized IC* against the reference data. The model comparison with the Bayes factor suggests strong evidence in favor of the *Generalized IC*, as expected and similar to the rectangular inclusions.

6.4 Summary and conclusions

We have applied the surrogate-assisted uncertainty-aware Bayesian validation framework to a benchmark study that addresses both parametric and conceptual modeling uncertainties due to different formulations of coupling free flow and porous medium models. To do so, we have considered the Stokes equations coupled to different models for the porous-medium compartment and corresponding coupling strategies: the standard REV-scale model using Darcy's law with classical or generalized interface conditions as well as the pore-network model. The advantage of employing a surrogate modeling technique is that one can perform a sensitivity analysis without additional costs. This analysis is achieved using the so-called Sobol indices that are derived analytically from the expansion coefficients.

Applying the suggested Bayesian validation framework, we have observed that there are matches between the predictions related to the considered models and the reference data for the points in the deeper part of the porous medium for all coupled models. However, we have found differences in the predictive capabilities of the models in the vicinity of the interface and in the free-flow region. Moreover, we have propagated the post-calibration parametric uncertainty through each analyzed model to validate the different models against reference data that have not been used during the calibration phase. This uncertainty-aware Bayesian validation procedure confirmed that the averaged pore-network model has the most difficulties correctly representing the underlying physical process. This issue is likely due to the averaging approach used for the pore-network model, where velocities have to be calculated and interpolated from fluxes only given within pore throats. Addressing the differences in the predictions of the considered modeling concepts, we have performed a Bayesian model comparison. This comparison revealed that the Stokes–Darcy model with the generalized interface conditions represents processes on the REV scale best compared to the classical interface conditions and the correspondingly upscaled pore-network model. The pore-network model outperforms both Stokes–Darcy models with classical and generalized interface conditions only if the detailed pore-scale information is considered.

We have also investigated two other cases: one with the opening boundary condition on the left side and another with circular inclusions in the porous medium. The analysis of the former setting, which induces parallel flow to the interface, uncovered that the Stokes–Darcy models with the classical and the generalized interface conditions provide similar results. This observation was expected for flows parallel to the fluid–porous interface. The findings of the analysis for the setup with circular inclusions confirmed that there is decisive evidence in favor of the generalized interface condition being superior to the classical interface. Concluding, we have observed that the suggested surrogate-assisted uncertainty-aware Bayesian validation framework helps to gain insight into underlying physical processes at considerably low computational costs.

7 Application III: Surrogate-Based Bayesian Comparison of Computationally Expensive Models: Application to Microbially-Induced Calcite Precipitation

Geochemical processes change the material properties of porous media in subsurface reservoirs caused by microbial activity. These complex Subsurface biogeochemical processes are subject to strong conceptual uncertainty at present. That means modelers face the challenge of choosing one of several modeling approaches to describe the biogeochemical process. Different hypotheses about the involved governing processes are present in these model variants.

When observation data are available, a rigorous Bayesian model selection in conjunction with a Bayesian model justifiability analysis can be employed to select the most appropriate model, which best describes the underlying physical processes in the light of the available data. However, biogeochemical modeling is computationally demanding as it conceptualizes different phases, biomass dynamics, geochemistry, precipitation, and dissolution in porous media. As a result, the two-stage Bayesian framework for the multi-model comparison cannot be based directly on the full computational models due to the high computational costs involved. To circumvent this problem, the surrogate-based framework introduced in Section 2.3 has been deployed for the competing biogeochemical models. Given that those surrogate representations are only approximations of the analyzed original models, the approximation error in the Bayesian analysis is accounted for by novel correction factors for the resulting model weights, introduced

in Sections 2.3.1 and 2.3.2. This work has been published in the Journal of Computational Geosciences titled "Surrogate-based Bayesian Comparison of Computationally Expensive Models: Application to Microbially Induced Calcite Precipitation" [Scheurer et al., 2021].

7.1 Introduction

Porous media undergo biogeochemical changes due to the activity of microbes [Lovley and Chapelle, 1995]. Due to the presence of these activities in the subsurface, they profoundly affect ecosystems, which makes them an interesting engineering application. Some examples of biogeochemical processes that engineers tried to manipulate are: enhanced recovery of resources as in microbially enhanced oil recovery, [e.g., Bachmann et al., 2014, McInerney et al., 2005, Huang et al., 2018], blocking of preferential flow paths by the accumulation of biomass or minerals precipitated as a result of the microbial metabolism, [e.g., Bottero et al., 2013, Suliman et al., 2006] or bioremediation of soils by microbial decomposition of organic pollutants, [e.g., Megharaj et al., 2011, Head, 1998, Mulligan and Galvez-Cloutier, 2003] or in-situ sequestration of inorganic contaminants (metals, radionuclides) by biotically managed precipitation [Hamdan et al., 2011].

Nevertheless, it is challenging to describe all biogeochemical processes in detail due to their complex interactions [Steefel and MacQuarrie, 1996]. As a result, it is hard to control them in a desired manner. 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 [Hunter et al., 1998]. Corresponding models are essential in investigating the coupled transport of fluids and reactive substances through porous media and the resulting chemical reactions in the pores [Steefel et al., 2005, MacQuarrie and Mayer, 2005, Xu et al., 2006].

Several transport models dealing with the biogeochemical process of microbially induced calcite precipitation (MICP) have been discussed in works by, e.g., Barkouki et al. [2011], Ebigbo et al. [2012], Hommel et al. [2015, 2016], van Wijngaarden et al. [2016], Nassar et al. [2018]. 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., Stocks-Fischer et al., 1999, Dupraz et al., 2009, Phillips et al., 2013, Cuthbert et al., 2013, Mitchell et al., 2013]. MICP has also been proven to reduce permeability and enhance mechanical strength even at large, field-relevant scales [e.g., van Paassen et al., 2010, Phillips et al., 2016, Nassar et al., 2018, Minto et al., 2019, Kirkland et al., 2020]. Additionally, MICP can be used to reduce erosion or increase soil stability [e.g., Whiffin et al., 2007, Gomez et al., 2017, van Paassen et al., 2010, Yang et al., 2020].

In porous media, MICP involves several phases: at least three solid phases (biofilm, calcite, unreactive solid substance), water, and possibly another fluid phase, such as gas. Moreover, calcium, inorganic carbon, and urea are considered dissolved components in the water phase. The complete list of components can be found in Hommel et al. [2015]. 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. The MICP processes are visualized in Figure 7.1.

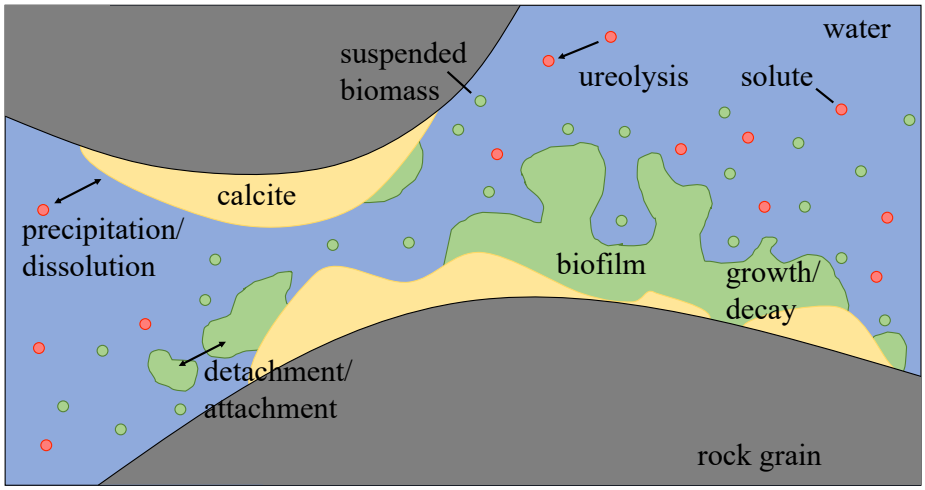


Figure 7.1: Schematic view of relevant processes and phases during MICP after Hommel et al. [2015].

S. pasteurii are bacteria that are able to produce the enzyme urease and decompose

urea into carbonic acid and ammonia with the help of urease. In an 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 reacts with the carbonate ions and calcite precipitates. These chemical processes can be described by the following reaction according to Hommel et al. [2015]:

$$CO(NH_2)_2 + 2\,H_2O + Ca^{2+} \xrightarrow{\text{Urease}} 2\,{NH_4}^+ + CaCO_3 \downarrow . \tag{7.1}$$

Conceptual uncertainty Biogeochemical models are helpful, for example, to design, monitor, and evaluate applications, such as mitigating leakages from a geological gas reservoir into above aquifers in advance, [e.g., Cuthbert et al., 2013, Nassar et al., 2018, Cunningham et al., 2019, Minto et al., 2019, Landa-Marbán et al., 2021]. Our limited knowledge about the interaction of the governing processes in biogeochemical systems leads to several modeling approaches. These approaches differ, e.g., in their level of complexity. Conceptual uncertainty refers to the uncertainty of selecting between these modeling alternatives.

Selecting a single model and ignoring possible alternatives might result in a substantial underestimation of the overall prediction uncertainty as the space of potential models is not sufficiently covered [Enemark et al., 2019, Refsgaard et al., 2012, Rojas et al., 2008]. Many studies have identified conceptual uncertainty as a key source of uncertainty in modeling [Burnham and Anderson, 2002, Neuman, 2003, Højberg and Refsgaard, 2005, Rojas et al., 2008, 2010, Gupta et al., 2012, Troldborg et al., 2007, Refsgaard et al., 2012, Renard et al., 2010, Schöniger et al., 2015b, Enemark et al., 2019]. These studies suggest treating modeling concepts with different levels of detail and different assumptions as competing hypotheses. A statistical technique such as BMS (Section 2.2.2) allows us to determine the system's most appropriate representation [Raftery, 1995, Wasserman, 2000].

This study aims to set up a rigorous ranking of biogeochemical computationally expensive models via the surrogate-based two-stage Bayesian multi-model comparison framework, introduced in Section 2.3. This framework is an extension of the Bayesian model justifiability analysis introduced by Schöniger et al. [2015a]. This extension allows the use of surrogate models, making this analysis suitable for computationally

demanding models. Section 7.2 introduces the experimental setup and the modeling variants. The results and discussions are provided in Section 7.3.

7.2 Problem description

This section entails the description of the experimental setup as well as the investigated numerical models. Moreover, the solution procedure used to perform the analysis is presented.

7.2.1 Experimental setup

The analyzed MICP experiment is described in detail in Hommel et al. [2015]. This experiment represents a sand-filled column that is 61 cm high with a diameter of 2.54 cm. At the beginning of the experiment, bacteria are injected at the bottom of the column. In an overnight no-flow period, bacteria can attach throughout the column and establish a biofilm. 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 the substrate to revive the biofilm [Hommel et al., 2015], before the subsequent injections of calcium and urea start over until a total number of 30 cycles has reached. A schematic experiment setup is shown in Figure 7.2.

This study considered the model predictions of calcium and calcite over space and time among various predicted quantities. The predictions of different models are compared to measurement data as well as among each other. To receive comparable results, only spatial and temporal points where measurement data are available are used when comparing models with each other. These data points differ for calcium and calcite. For the calcite content, the 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

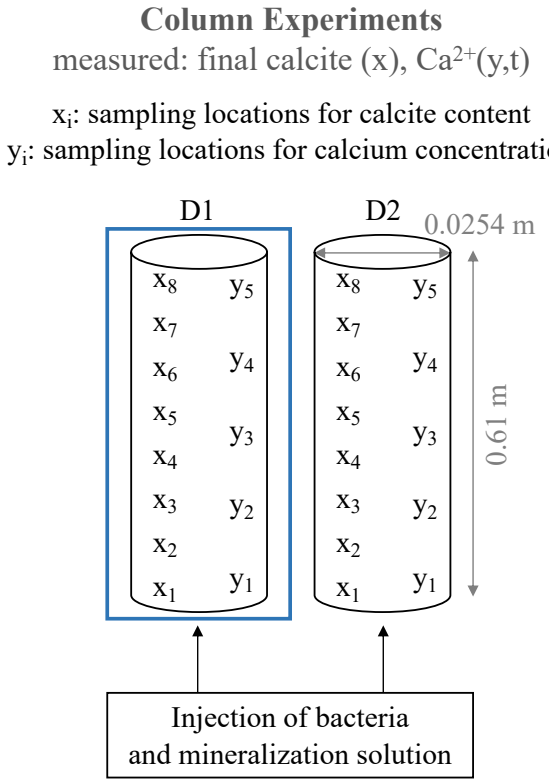


Figure 7.2: Column experiment setup by Hommel et al. [2015] with measurement locations for calcite content and calcium concentration with analyzed column D1.

Table 7.1: Times in hours for measurement of the calcium concentration.

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

concentration is measured additionally after half an hour, one, two, three, and four hours, except for pulse 22, where no measurement is available after 3 hours, resulting in 35 temporal points. Table 7.1 provides the exact measurement times after the first injection. 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 49.53 cm distance from the bottom. The measurement locations in the models are evenly distributed at a respective distance of half an inch (1.27 cm).

7.2.2 Conceptual models and related uncertainty

This study analyzed three models of MICP for describing biogeochemical processes in porous media, provided by Hommel et al. [2015, 2016]. The reader is referred to the original publications for a detailed explanation of their equations and the considered numerical schemes. 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 a fully-coupled-vertex-centered finite volume (box) scheme [Helmig, 1997] in space; the system of equations is solved using the BiCGStab solver [Van der Vorst, 1992] 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 depends on the model parameter set chosen for the evaluation since the time stepping varies adaptively. Therefore, Hommel et al. [2015] suggest two simplifications of the *full complexity* model, M_{FC}, using the certain 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 Equation 7.1 [Hommel et al., 2016].

Table 7.2 summarizes the key differences relevant to the model simplifications

Table 7.2: Key differences of the investigated models.

model	M_{FC}	M_{IB}	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

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 obtaining results of one model run after 40 minutes using the same computational cluster. Apart from decreasing the computational cost, model simplification reduces parametric uncertainty. An extremely 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.

The considered parameters in the following were previously identified as sensitive parameters of the MICP models and already used for calibration in Hommel et al. [2015]:

- the coefficient for preferential attachment to biomass $c_{\mathrm{a},1}$,
- the coefficient for attachment to arbitrary surfaces $c_{\mathrm{a},2}$,
- the dry mass density of biofilm ρ_{f},
- the enzyme content of biomass k_{ub}.

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 7.3.

Table 7.3: Uncertain parameters and their prior distributions for the MICP models.

Parameter name	Range	Unit	Distribution type
$c_{\mathrm{a},1}$	$[1 \cdot 10^{-10}, 1 \cdot 10^{-7}]$	s^{-1}	uniform
$c_{\mathrm{a},2}$	$[1 \cdot 10^{-10}, 1 \cdot 10^{-6}]$	s^{-1}	uniform
ρ_{f}	$[1, 15]$	$\mathrm{kg/m^3}$	uniform
k_{ub}	$[1 \cdot 10^{-5}, 5 \cdot 10^{-4}]$	kg/kg	uniform

7.2.3 Solution procedure

Two surrogate models (one for calcite, one for calcium) have been trained for each of the three competing MICP models described in Section 7.2.2 (resulting in a total of six different surrogate models) using a $p = 2$ order aPC expansion according to the prior distributions presented in Table 7.3. For this purpose, the three original models will be evaluated $D = card\ \mathcal{A}^{d,p} = (d+p)!/(d! \cdot p!)$ times. Since the D evaluations for constructing the surrogate models are independent, these model runs were parallelized. Further, we improve the accuracy of the three surrogates using an iterative Bayesian updating method by incorporating new collocation points at approximate locations of the maximum *a posteriori* parameter set [Oladyshkin et al., 2013, Mohammadi et al., 2018]. The idea is to evaluate the surrogate model $\tilde{M}$ on a high number of parameter

realizations obtained from their prior distribution, to assign weights to the points by their posterior probability, and select the parameter set with the highest weight.

The number of Bayesian updates have been restricted to 10 runs due to the high computational demand and previous experience (see e.g. Beckers et al. [2020]), so that $D_{\text{total}} = D + 10 = (d+p)!/(d! \cdot p!) + 10$. This results in $D_{\text{total}} = 15 + 10 = 25$ model evaluations for the *simple chemistry* model (M_{SC}) and the *full complexity* model (M_{FC}) and $D_{\text{total}} = 6+10 = 16$ 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.

7.3 Results and discussion

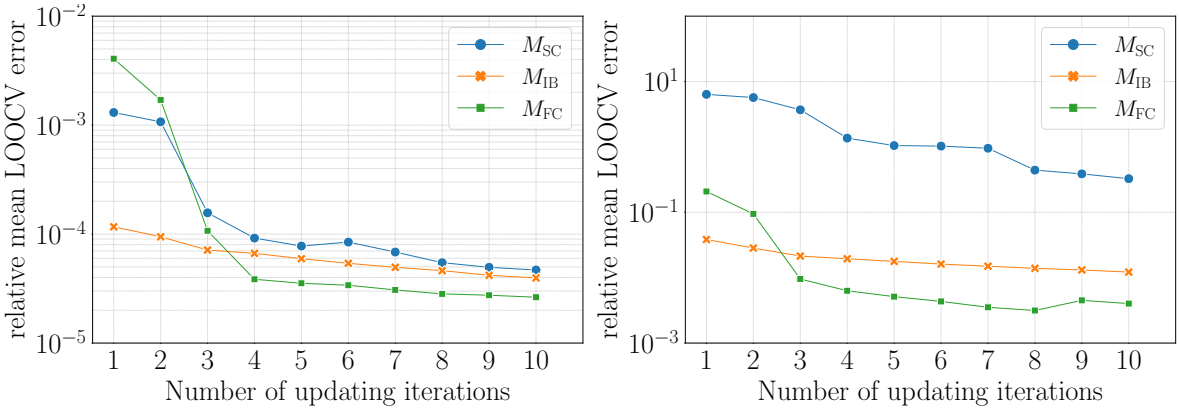
7.3.1 Approximation quality of MICP surrogate models

When iteratively updating the aPC expansion, a surrogate model's quality can be assessed by estimating its validation error in Equation (3.37). This can be computed for a validation/test set as discussed in Section 3.1.3. Since the computation of the validation error requires additional model evaluations, it is computationally intractable for models with significant computational time, such as the MICP models. To avoid additional model evaluations for assessing the accuracy of the surrogate model, the LOO error in Equation (3.38) based on the already evaluated training model runs can be used.

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, the respective error values have been averaged over space (and time) after every updating step. Moreover, since the two quantities of interest (calcite content [%] and calcium concentration [mol/m^3])

are in different orders of magnitude, the relative errors of the surrogate models for calcium and calcite need to be computed by normalizing to the mean output value.

Figure 7.3 provides the errors for all considered models for calcite content (left) and calcium concentration (right) during the Bayesian updating process. 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.



(a) Calcite content (b) Calcium concentration

Figure 7.3: Relative mean LOOCV errors for SRQs during Bayesian updating.

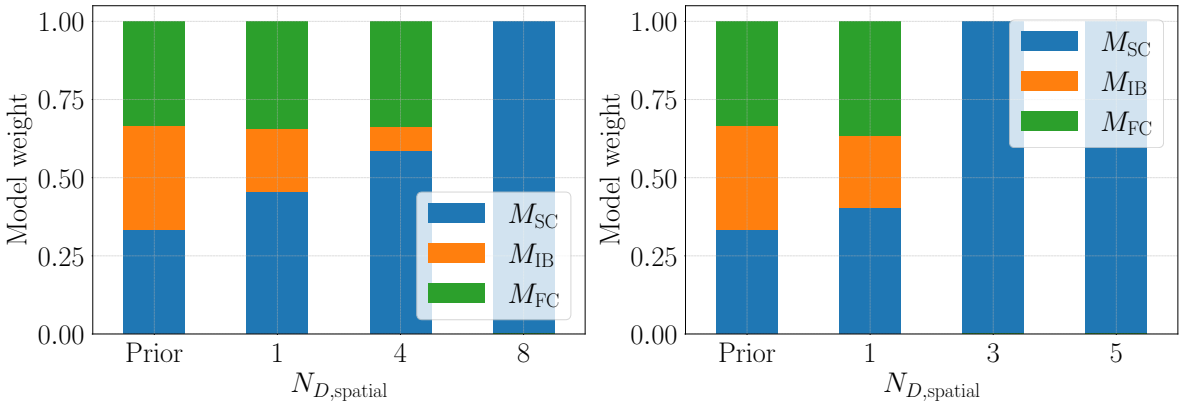
The figure reveals that the error for calcite decreases more strongly than that of calcium. It is also remarkable that the error for calcite is in a similar order of magnitude for all models. This means that all surrogate models are of comparable quality for the calcite content. For calcium, the error of the *simple chemistry* model M_{SC} is significantly larger than that of the other two surrogate models. This can occur if one uses Bayesian updating and seeks 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. Still, it deviates greatly 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 has been compensated in the model weight calculations by the newly introduced correction factor, as presented in Section 2.3.

7.3.2 Two-stage Bayesian multi-model comparison

We have performed the two-stage surrogate-based Bayesian multi-model comparison, consisting of a model selection incorporating the measurement data and a Bayesian model justifiability analysis according to Section 2.3 using the trained surrogate representations of the three analyzed MICP models from Section 7.3.1. Following the justifiability analysis, we compute the model weights as stated in Section 2.2.2 and adjust them with the novel correction factors, introduced in Section 2.3.2.

BME convergence was ensured by checking the average likelihood's evolution over an increasing data set size. To justify the underlying physical assumptions behind the MICP models, we have assessed the impact of the data set size on BME values appearing in the Bayesian model justifiability analysis. This has been achieved by starting with only one spatial data point, then using half of the available data set size, and finally including all the spatial data points for calcium and calcite. This division results in the 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.

Bayesian model selection The results of the BMS analysis based on the measurement data is presented in Figure 7.4.



(a) Calcite content (b) Calcium concentration

Figure 7.4: Model weights for the prediction of SRQs over the increasing amount of used spatial data points $N_{D,spatial}$.

The model weights suggest that the *simple chemistry* model M_{SC} obtains the highest model weight (normalized BME value) for all data set sizes. A model wins the com-

petition either because of its low complexity or because of its goodness-of-fit to the measurement data (or both) Schöniger et al. [2015a]. The justifiability analysis will further investigate these two aspects in the second stage.

Model justifiability analysis Figure 7.5 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.

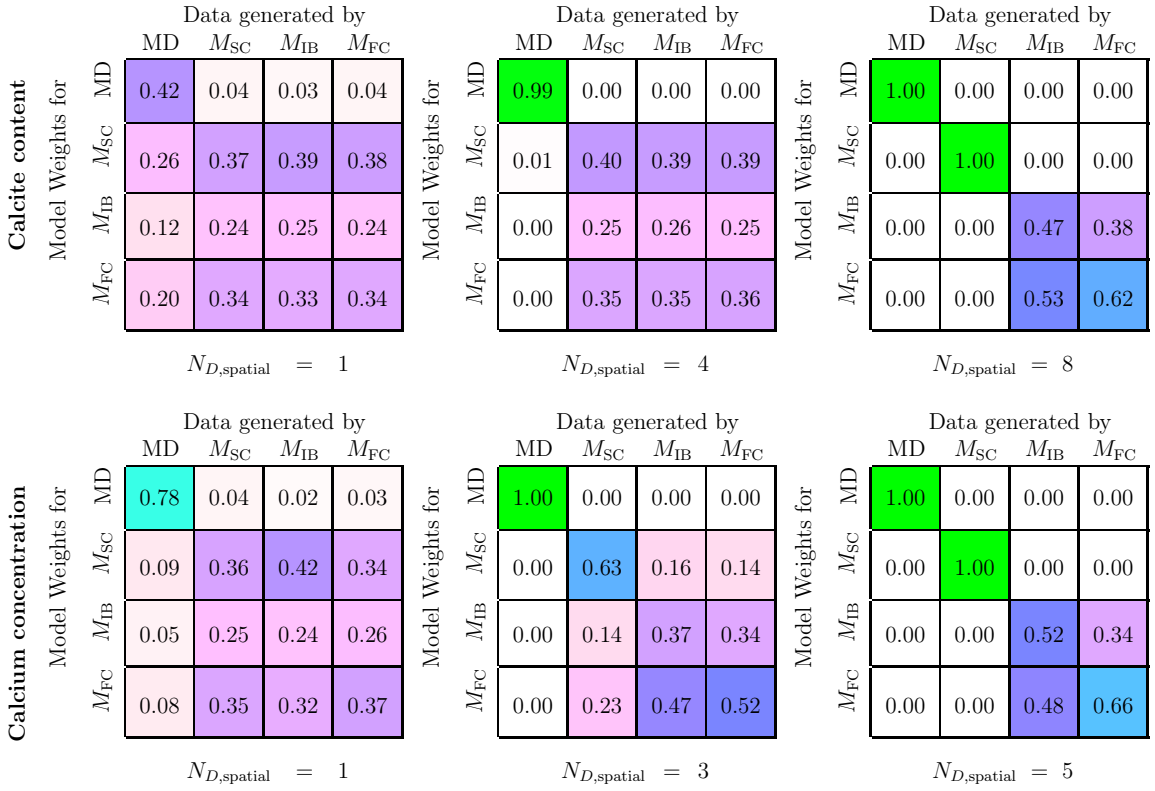


Figure 7.5: Model confusion matrices for calcite content [%] and calcium concentration [mol/m^3] of the three models and the measurement data (MD) over increasing amount of used spatial data points $N_{D,\text{spatial}}$.

The main-diagonal entries of the model confusion matrices in Figure 7.5 represent the models' ability to identify their own predictions. The higher the value of the main diagonal entry, the higher is the probability of the model identifying itself as the data-generating process. The diagonal values increase when a larger data set size is used,

agreeing well with the theory of the Bayesian model justifiability analysis discussed in Schöniger et al. [2015a].

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 entire 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 reveals that the *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 extremely low. In statistical terms, this means that all models are clearly rejected by the full data set. This is in accordance with the conclusions drawn in Hommel et al. [2015], that there is at least one relevant process not yet implemented in "sufficient detail", which is necessary for better results.

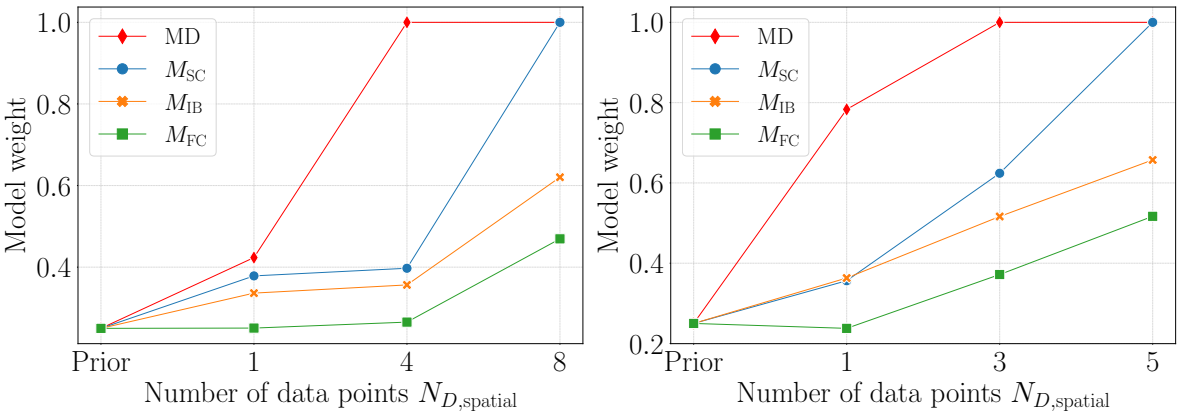
How much data is needed?

The confusion matrices on the left in Figure 7.5 show that considering only one spatial data point is insufficient since the model weights belonging to 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 justification for any model and even the measurement data of the calcite content are not able to identify itself (which is evident 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, corresponding to the principle of parsimony.

When using half of the data set shown in Figure 7.5, the second column, 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 M_{IB}'s weight for the calcium concentration reaches the absolute majority, which means that the data set size is sufficient for self-identification. As a result, the physical model assumptions leading to simplifications are justifiable.

Figure 7.6 reveals the change of self-identification weights (the confusion matrix's main diagonal entries) over an increasing data set size. According to the figure, perfect



(a) Calcite content (b) Calcium concentration

Figure 7.6: Average model weights for the data-generating process of the two SRQs of the three models and the measurement data (MD) over the increasing amount of used spatial data points $N_{D,spatial}$.

justification (model weight of 1.00) is achieved very quickly with the simplest model M_{SC} and clearly for the measurement data. For the *initial biofilm* model M_{IB} and the *full complexity* model M_{FC}, a larger data set size is needed to justify their complexity. As the weights for the more complex models do not stagnate at some point, it is not expected that a much larger data set could necessarily 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.

How similar are the models?

To assess the model similarities, one needs to explore the off-diagonal entries of the confusion matrix in Figure 7.5. For a single data point, we can clearly see that the models "confuse" their predictions, as the off-diagonal weights are relatively high. For the case in which 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 competing 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.

Contrary to that, the model confusion remains for the calcium predictions, i.e., the current data set size does not yield a more apparent 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. Considering only the calcite content reveals 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, according to Schöniger et al. [2015a].

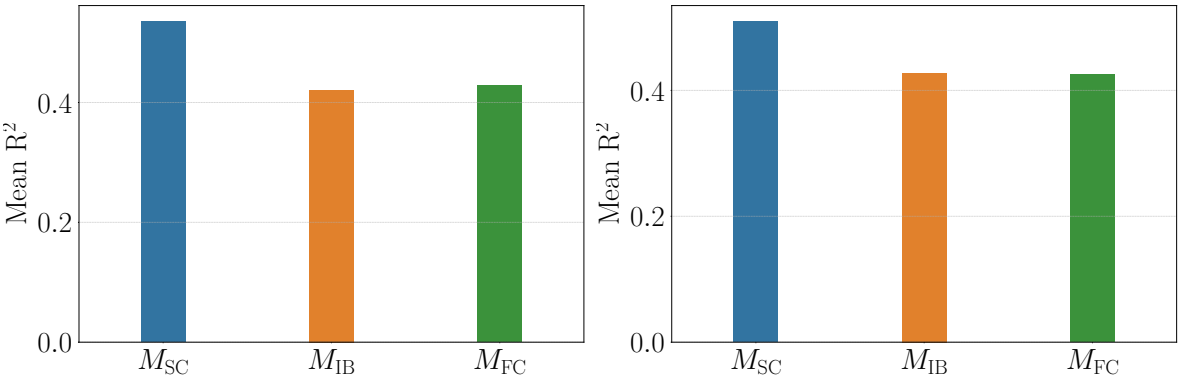
How do models fit the data?

The goodness-of-fit of the models to the measurement data can be investigated via the coefficient of determination (R^2) between the different model outputs and the corresponding measurement data, averaged over all model runs evaluated on D_{total} different surrogate training points, which reads

$$R^2 = \frac{1}{D_{total}} \sum_{i=1}^{D_{total}} \left(\frac{\sum_{j=1}^{N_{\mathrm{s}}} \left(M_{k,j}\left(\boldsymbol{\theta}^{(i)}\right) - \overline{\boldsymbol{y}}_0 \right)^2}{\sum_{j=1}^{N_{\mathrm{s}}} \left(\boldsymbol{y}_{0,j} - \overline{\boldsymbol{y}}_0 \right)^2} \right). \tag{7.2}$$

$\boldsymbol{y}_{0,j}$ is the vector of measurements at position j of total length N_s and its mean is denoted by $\overline{\boldsymbol{y}}_0$. Moreover, $M_{k,j}\left(\boldsymbol{\theta}^{(i)}\right)$ represents the model output of model M_k at position j evaluated at collocation point $\boldsymbol{\theta}^{(i)}$.

Figure 7.7 displays the goodness-of-fit to the measurement data for the calcite content and the calcium concentration. 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, the calcite content and the calcium concentration predictions, *simple chemistry* model M_SC shows the highest mean R^2. 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. Note that a more complex model requires a significantly better goodness-of-fit to justify its complexity [Schöniger et al., 2015a] (and to achieve a similar weight as a simpler model). Furthermore, it is remarkable 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.



(a) Calcite content (b) Calcium concentration

Figure 7.7: Mean R^2 between the different model outputs and the measurement data.

7.4 Summary and conclusions

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. The

study presented in this chapter applied the surrogate-based Bayesian model justifiability analysis to analyze microbially induced calcite precipitation models in porous media. The suggested framework offers a rigorous pathway to address the so-called conceptual uncertainty, i.e., which model best describes the underlying physical system. The justifiability analysis compares the models with each other and the available measurement data.

Applying the justifiability analysis in addition to the BMS analysis yields a better insight into 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 [Schöniger et al., 2015a].

Applying the 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 calcite content and calcium concentration predictions 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.

Combining the insights from the Bayesian model justifiability analysis and the goodness-of-fit analysis, the following conclusions about the *initial biofilm* model M_{IB} and *simple chemistry* model M_{SC} as simplifications of the *full complexity* model M_{FC} can be drawn:

- 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.
- The M_{IB} model 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 (Figure 7.6) and Bayesian model justifiability (Figure 7.5) with the full data set.
- Therefore, the *simple chemistry* model M_{SC} uses its full potential to represent the data and appropriately captures the response of the underlying physical system.

The proposed surrogate-based justifiability analysis is an extension of the general justifiability analysis proposed by Schöniger et al. [2015a], making it applicable for computationally expensive models. However, with increasing computational cost and limited computational budget, i.e., the number of model evaluations, the justifiability analysis' results become less reliable, as the surrogates provide imprecise predictions.

8 Summary and Outlook

8.1 Summary

Given the importance of computational modeling in geoscience, especially in porous media research, assessment of the quality of models in light of the purpose of a given simulation is of paramount importance to engineering designers and managers, public officials, and those affected by the decisions based on the predictions. One of these assessment tools is validation, which investigates how accurately a computational model describes reality. Graphical comparisons of computational results and experimental data are common practice in many fields. These graphical comparisons are usually made by plotting some computational system response quantity with the experimentally measured response over a range of input parameters. If the computational results generally agree with the experimental data, the computational model is commonly declared, "validated". With this graphical comparison, the effects of uncertainty in the computational results and experimental data are not typically quoted, nor is its statistical character quantified.

To address the abovementioned issues, I presented a statistical framework to establish quantitative uncertainty-aware validation benchmarks. In the multi-model comparison setting, models are analyzed in a validation benchmark via a statistical model selection approach called Bayesian Model Selection (BMS). In BMS, model probabilities are used to express uncertainty between models regarding how likely a candidate model is to generate the observed data. BMS also allows for determining the impact of the data size in the ranking outcome and the similarity of the models or lack thereof. The first issue answers the question: how much data are required to achieve reasonable ranking or which level of complexity is justified given the available data? However, the similarity can be addressed by the question: how would the models be ranked given

that one of them is the data-generating problem? These aspects were included in a so-called justifiability analysis introduced by Schöniger et al. [2015a].

This framework also utilizes Bayesian hypothesis testing (BHT) as a tool for statistical decision-making of whether a computational model is superior to this counterpart. BHT incorporates the use of the Bayes factor to evaluate the plausibility of rejecting a null hypothesis according to a designated confidence level. The Bayes factor is interpreted as the ratio of the relative likelihood of the null hypothesis over the alternative hypothesis. The null hypothesis states that the experimental data support the predictions of one model, while the alternative hypothesis states that the competing model includes the data-generating processes.

The probabilistic nature of the Bayesian techniques, such as BHT and BMS used in the framework, requires propagating the parametric uncertainty through all competing models – i.e., a significant number of model evaluations – to reach statistical convergence. In practice, however, the computational complexity of the underlying computational model and the total available computational budget severely restrict the number of evaluations one can perform. By replacing computational models with easy-to-evaluate surrogates, one can mimic the behavior of underlying physical models using a limited set of runs without sacrificing accuracy.

The developed methodology aims to help modelers perform uncertainty-aware model validation benchmarks. A two-stage Bayesian multi-model framework is discussed for modeling tasks where a set of models are at hand. To make this framework applicable for computationally demanding models, it is extended to a surrogate-assisted framework, which keeps the computational costs at a reasonable level. I employed polynomial chaos expansion (PCE) to build a surrogate representation of models. Moreover, correction factors were introduced to compensate for the surrogate error in the BHS and BMS, as using surrogate representations instead of the full-fidelity models introduces additional errors to the validation metrics.

In this dissertation, I showed how Bayesian formalism could be materialized by employing the concept of PCE to achieve more accurate surrogates with a sparse PCE representation and account for the uncertainty in the surrogate's predictions. I also highlighted how a surrogate model using a PCE could be constructed with as few simulations as the computational budget allows. I discussed sequential adaptive sampling strategies, in which one attempts to augment the initial design iteratively. By doing

so, informative regions in the parameter space are adequately explored. These regions are more likely to provide valuable information on the behavior of the original model responses. Using a sequential sampling strategy avoids the waste of computational resources, as opposed to the so-called one-shot designs. A series of benchmark studies were conducted to investigate the predictive capabilities of different sparsity and sequential adaptive sampling methods.

Moreover, I introduced *BayesValidRox*, an open-source, object-oriented Python package that provides an automated workflow for surrogate-based sensitivity analysis, Bayesian calibration, and validation of computational models with a modular structure. The uncertainty-aware validation framework was applied to a range of cases in the field of subsurface hydro-system modeling, mainly to flow and transport in porous media, such as flow simulation models in fractured porous media, coupling free flow and porous medium flow, and microbially induced calcite precipitation. However, this validation framework can be transferred to other disciplines in which models are used, such as psychology, ecology, economics, and other engineering disciplines.

8.2 Outlook

Using the insights and conclusions from this dissertation, one can explore new avenues in future research. Regarding surrogate modeling, the following aspects can be investigated to improve the performance of the framework:

Surrogate modeling with dependent inputs The traditional PCE can only deal with independent parameters. However, in my engineering applications, parameter dependency does exist in computational models. The framework can be extended to handle the models with interdependent parameters. One approach is to use a proper function to transform the distribution of the parameters into independent standard normal distributions leading to independent orthogonal polynomial bases. The common transformation functions are Nataf, Box-Cox, Rosenblatt, exponential, and logarithmic. Another approach proposed by Ghaith et al. [2021] is to leverage the potential of principal component analysis for addressing the parameter dependency. Their proposed PCA-PCE framework aims to decouple the correlations among dependent parameters

via principal component analysis (PCA) and to generate independent standard normal distributions using Johnson distributions for further development of the PCE approach.

Surrogate modeling for time-dependent quantities In Section 3.1.5, I briefly discussed a data compression approach to handle time-dependent quantities. Instead of applying PCEs at each time instant, a dimensionality reduction is used to capture the main stochastic features of the response quantities by a small number of the principal components. Traditionally, dimensionality reduction is performed using a linear technique such as Principal Components Analysis. However, these linear techniques cannot adequately handle complex nonlinear responses. To address this issue, nonlinear dimension reductions can be employed to deal with complex nonlinear model responses. For an overview of these techniques, see, e.g., Saul et al. [2006], Lee and Verleysen [2007], Venna et al. [2007], Van Der Maaten et al. [2009], Burges [2010].

As principal components increase in order, PCE accuracy reduces [Blatman and Sudret, 2013]. Vohra et al. [2020] argue that this is because not all input parameters are required for mapping the input space to the principal components. They proposed a framework that employs the so-called active spaces [Constantine, 2015] with typically fewer dimensions to map to principal components. The combination of this approach with a PCA-PCE framework could be a promising future direction as far as the surrogate modeling of time-dependent quantities is concerned.

A stochastic time-warping method is another option for handling time-dependent responses. All realizations of the model are rescaled to a common time scale so that they appear in-phase in a virtual time stamp. In other words, this method increases the similarities among the realizations in a preprocessing step so that responses vary in a small neighborhood of a reference realization. Eventually, PCEs are applied to rescaled realizations, which have lower variability. For more details see Mai and Sudret [2017] and the references therein.

A model-time allocation strategy The proposed framework in this dissertation distributes the computer power evenly in a multi-model setting. This even distribution means that the number of model evaluations is equal for all models. This approach may lead to the waste of computational budgets in a benchmarking task when models have different properties, such as the dimension of parameter spaces, smoothness, prior

probability, and computing time. Since these factors are typically not known before any evaluations, one can think of a clever way to allocate the computational resources to the right model in each refinement iteration. Sinsbeck et al. [2021] proposed to equip the sequential adaptive experimental design (SAED) with the model-time allocation methods. They consider what is known about the models in each SAED iteration and then run the model that contributes the largest expected gain in accuracy.

As for the other parts of the proposed framework, future research shall address the following aspects:

Statistical model discrepancy The model discrepancy always exists for various reasons, such as simplified assumptions, missing physics, and upscaling due to scale differences. Several methods have been proposed in the literature to incorporate the model discrepancy in a Bayesian setting. These methods' treatment of model discrepancy range from a constant bias to more sophisticated statistical methods in which one forms a Bayesian hierarchical model to solve a joint parameter and model discrepancy inference problem [Kennedy and O'Hagan, 2001, Bayarri et al., 2007, Brynjarsdóttir and O'Hagan, 2014, Ling et al., 2014, Gardner et al., 2021]. Statistical error models for each process-based model have not been explicitly included within BMS to reduce prediction bias. It is expected that incorporating model discrepancy into the BMS routine will further improve the predictive coverage and allow a potentially more realistic estimation of model parameters and their associated uncertainties.

Validation metrics In many engineering applications, there is no possibility to replicate the measurements many times to estimate a measurement uncertainty to be included in the Bayesian setting. In these cases, it is common to assume a measurement error. However, this assumption could lead to misleading results. Therefore, more suitable validation metrics that could measure the difference between a deterministic measurement value and a predictive distribution for the stated cases would generate more reliable results and reduce the degree of subjectivity in the validation process. One metric is the so-called *energy score* [Gneiting and Raftery, 2007], which is the deterministic counterpart of the energy distance [Rizzo and Székely, 2016]. Schäfer Rodrigues Silva et al. [2022] successfully employed the energy score for diagnosing similarities in probabilistic multi-model ensembles: an application to soil–plant-growth-modeling.

A MCMC Convergence

I use two convergence methods: Robin & Gelman method and integrated autocorrelation time, whose definitions are to follow.

Gelman and Rubin [1992] introduce a quantitative approach to assess the convergence of an MCMC sampler. Their approach compares a set of L independent Markov chains, i.e., walkers, initialized at different seed points. The chains are believed to converge when the empirical second moments (variance) computed from the individual chains are the same as that of combined samples from all L walkers. Both variances are combined in a weighted sum to estimate the variance of a parameter θ.

Let $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_L\}$ be the L parallel walkers, starting from different seed points $\boldsymbol{\theta}_i^{(0)}$. Moreover, each chain $\boldsymbol{\theta}_i$ contains $K+1$ sample points $(\boldsymbol{\theta}_i^{(0)}, \cdots, \boldsymbol{\theta}_i^{(K)})$ with $\boldsymbol{\theta}_i^{(k)} \in \mathbb{R}^M$. The covariance matrix of the i-th walker is estimated by

$$\boldsymbol{W}_i = \frac{1}{K} \sum_{k=0}^{K} \left(\boldsymbol{\theta}_i^{(k)} - \boldsymbol{\mu}_i\right) \left(\boldsymbol{\theta}_i^{(k)} - \boldsymbol{\mu}_i\right)^\top, \quad \boldsymbol{\mu}_i = \frac{1}{K+1} \sum_{k=0}^{K} \boldsymbol{\theta}_i^{(k)}. \tag{A.1}$$

To compute the within-chain covariance $\boldsymbol{W}$, the covariance matrices for all L walkers are averaged

$$\boldsymbol{W} = \frac{1}{K} \sum_{i=1}^{L} \boldsymbol{W}_i. \tag{A.2}$$

The so-called between-chain variance $\boldsymbol{B}$, which captures the covariance between the individual chains, are computed by

$$B = \frac{K}{L-1} \sum_{i=1}^{L} \left(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}\right) \left(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}\right)^\top \tag{A.3}$$

where

$$\overline{\boldsymbol{\mu}} = \frac{1}{L(K+1)} \sum_{i=1}^{L} \sum_{k=0}^{K} \boldsymbol{\theta}_i^{(k)}. \tag{A.4}$$

The estimated variance of $\boldsymbol{\theta}$ as the weighted sum of between and within chain variance can be computed by

$$\text{var}(\boldsymbol{\theta}) = \left(1 - \frac{1}{L}\right) W + \frac{1}{L} B \tag{A.5}$$

The potential scale reduction factor (PSRF) is computed by

$$\hat{R} = \sqrt{\frac{\text{var}(\boldsymbol{\theta})}{W}} \tag{A.6}$$

The metric in Equation (A.6) must be close to one for a well-converged chain. This condition indicates that the between-chain variance is small, which means chains mix around the stationary distribution. Gelman and Rubin [1992] show that $\hat{R}$ greater than 1.1 indicates that the chains have not yet fully converged and need a longer burn-in phase. In a burn-in phase, we discard sample points prior to convergence to avoid the pollution of the posterior estimation.

Brooks and Gelman [1998] later extended the PSRF method for multivariate cases with a set of parallel independent chains and proposed

$$\hat{R}^p = \frac{K}{K+1} + \left(\frac{L+1}{L}\right) \lambda_1, \tag{A.7}$$

where λ_1 represent the largest eigenvalue of the matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$, which is symmetric and positive definite. When $\hat{R}^p$ approaches one, the sampler's convergence is achievable.

Additionally, I monitor the convergence of the sampler using the so-called integrated autocorrelation time. This quantity estimates the number of evaluations of the posterior density function needed to draw independent samples from the target density. A more efficient chain is expected to have a shorter autocorrelation time. Sokal [1997] provides a great discussion of methods for autocorrelation estimation. An integrated

autocorrelation time $\tau_{\boldsymbol{\theta}}$ for an ensemble consisting of L chains can be defined as

$$\tau_\theta = \sum_{\tau=-\infty}^{\infty} \rho_\theta(\tau), \tag{A.8}$$

where $\rho_\theta(\tau)$ represents the normalized autocorrelation function of the stochastic process that generated the chain for θ. Using a finite chain $\{\theta_i\}_{i=1}^{L}$, one can approximate $\rho_\theta(\tau)$ via

$$\hat{\rho}_\theta(\tau) = \hat{C}_\theta(\tau)/\hat{C}_\theta(0) \tag{A.9}$$

where

$$\hat{C}_\theta(\tau) = \frac{1}{L-\tau} \sum_{i=1}^{L-\tau} \left(\theta^{(i)} - \mu_\theta\right)\left(\theta^{(i+\tau)} - \mu_\theta\right) \tag{A.10}$$

with μ_θ being the arithmetic average of the chains $\mu_\theta = 1/L \sum_{i=1}^{L} \theta_i$.

To estimate τ_θ, one can use

$$\hat{\tau}_\theta(P) = 1 + 2 \sum_{\tau=1}^{P} \hat{\rho}_\theta(\tau) \tag{A.11}$$

for some $P \ll L$. As discussed in Sokal [1997], using P chains of an ensemble instead of all L decreases the estimator's variance at the cost of some added bias. He also suggests choosing the smallest value of P where $P \geq C\hat{\tau}_\theta(P)$ with the constant C being 5. This procedure works well with chains longer than $50 \cdot \tau$. The python package *emcee* used for the framework provides an estimation of the integrated autocorrelation time. One can compute this measure for the chain of each parameter θ and use the maximum and minimum values $\tau_{\min} = \min_{0 \leq i < N} \hat{\tau}_{\boldsymbol{\theta}_j}$ and $\tau_{\max} = \max_{0 \leq i < N} \hat{\tau}_{\boldsymbol{\theta}_j}$. The sampler is run until the sample size $j > 100 \cdot \tau_{\max}$. The convergence of the MCMC sampler can be achieved if the difference in $\hat{\tau}_{\boldsymbol{\theta}}$ from sample $j - \tau_{\max}$ to sample j is less than 1 %.

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer New York, 1998.

R. Allison and J. Dunkley. Comparison of sampling techniques for Bayesian parameter estimation. *Monthly Notices of the Royal Astronomical Society*, 437(4):3918–3928, 12 2013.

D. Angluin and C. H. Smith. Inductive inference: Theory and methods. *ACM Computing Surveys (CSUR)*, 15(3):237–269, 1983.

P. Angot, B. Goyeau, and J. A. Ochoa-Tapia. Asymptotic modeling of transport phenomena at the interface between a fluid and a porous layer: Jump conditions. *Physical Review E*, 95(6):063302, 2017.

Y. Arjoune, N. Kaabouch, H. El Ghazi, and A. Tamtaoui. Compressive sensing: Performance comparison of sparse recovery algorithms. In *2017 IEEE 7th annual computing and communication workshop and conference (CCWC)*, pages 1–7. IEEE, 2017.

G. Aversano, J. C. Parra-Alvarez, B. J. Isaac, S. T. Smith, A. Coussement, O. Gicquel, and A. Parente. PCA and Kriging for the efficient exploration of consistency regions in uncertainty quantification. *Proceedings of the Combustion Institute*, 37(4):4461–4469, 2019.

S. D. Babacan, R. Molina, and A. K. Katsaggelos. Bayesian compressive sensing using laplace priors. *IEEE Transactions on image processing*, 19(1):53–63, 2009.

R. T. Bachmann, A. C. Johnson, and R. G. Edyvean. Biotechnology in the petroleum industry: an overview. *International Biodeterioration & Biodegradation*, 86:225–237, 2014.

O. Balci and R. G. Sargent. Some examples of simulation model validation using hypothesis testing. Technical report, Institute of Electrical and Electronics Engineers (IEEE), 1982.

M. T. Balhoff, K. E. Thompson, and M. Hjortsø. Coupling pore-scale networks to continuum-scale models of porous media. *Computers & Geosciences*, 33(3):393–410, 2007.

M. T. Balhoff, S. G. Thomas, and M. F. Wheeler. Mortar coupling and upscaling of pore-scale models. *Computational Geosciences*, 12(1):15–27, 2008.

T. Barkouki, B. Martinez, B. Mortensen, T. Weathers, J. De Jong, T. Ginn, N. Spycher, R. Smith, and Y. Fujita. Forward and inverse bio-geochemical modeling of microbially induced calcite precipitation in half-meter column experiments. *Transport in Porous Media*, 90(1):23–39, 2011.

M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, and J. Tu. A framework for validation of computer models. *Technometrics*, 49(2): 138–154, 2007.

H. Bazargan and M. Christie. Bayesian model selection for complex geological structures using polynomial chaos proxy. *Computational Geosciences*, 21(3):533–551, 2017.

H. Bazargan, M. Christie, A. H. Elsheikh, and M. Ahmadi. Surrogate accelerated sampling of reservoir models with complex structures using sparse polynomial chaos expansion. *Advances in Water Resources*, 86:385–399, 2015.

L. Beaude, K. Brenner, S. Lopez, R. Masson, and F. Smai. Non-isothermal compositional liquid gas darcy flow: formulation, soil-atmosphere boundary condition and application to high-energy geothermal simulations. *Computational Geosciences*, 23 (3):443–470, 2019.

G. S. Beavers and D. D. Joseph. Boundary conditions at a naturally permeable wall. *Journal of Fluid Mechanics*, 30(1):197–207, 1967.

J. Beck and S. Guillas. Sequential design with mutual information for computer experiments (mice): Emulation of a tsunami model. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):739–766, 2016.

F. Beckers, A. Heredia, M. Noack, W. Nowak, S. Wieprecht, and S. Oladyshkin. Bayesian calibration and validation of a large-scale and time-demanding sediment transport model. *Water Resources Research*, 56(7):e2019WR026966, 2020.

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

J. O. Berger and M. Delampady. Testing precise hypotheses. *Statistical Science*, 2(3): 317–335, 1987. ISSN 08834237.

J. O. Berger and J. Mortera. Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, 94(446):542–554, 1999.

J. O. Berger and T. Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397): 112–122, 1987.

B. Berkowitz. Characterizing flow and transport in fractured geological media: A review. *Advances in Water Resources*, 25(8-12):861–884, 2002.

J. M. Bernardo and A. F. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

I. Berre, F. Doster, and E. Keilegavlen. Flow in fractured porous media: a review of conceptual models and discretization approaches. *Transport in Porous Media*, 130 (1):215–236, 2019.

M. Berveiller, B. Sudret, and M. Lemaire. Presentation of two methods for computing the response coefficients in stochastic finite element analysis. In *Proc. 9th ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability, Albuquerque, USA*, 2004.

M. Berveiller, B. Sudret, and M. Lemaire. Stochastic finite element: a non intrusive approach by regression. *European Journal of Computational Mechanics/Revue Européenne de Mécanique Numérique*, 15(1-3):81–92, 2006.

S. Beyhaghi, Z. Xu, and K. M. Pillai. Achieving the inside–outside coupling during network simulation of isothermal drying of a porous medium in a turbulent flow. *Transport in Porous Media*, 114(3):823–842, 2016.

S. Bi, S. Prabhu, S. Cogan, and S. Atamturktur. Uncertainty quantification metrics with varying statistical information in model calibration and validation. *AIAA journal*, 55(10):3570–3583, 2017.

G. Blatman. *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis*. PhD thesis, Clermont-Ferrand 2, 2009.

G. Blatman and B. Sudret. Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of Computational Physics*, 230(6):2345–2367, 2011.

G. Blatman and B. Sudret. *Sparse polynomial chaos expansions of vector-valued response quantities*. CRC Press/Balkema, 2013.

M. J. Blunt. *Multiphase flow in permeable media: a pore-scale perspective*. Cambridge University Press, Cambridge, 2017.

S. Bottero, T. Storck, T. J. Heimovaara, M. C. van Loosdrecht, M. V. Enzien, and C. Picioreanu. Biofilm development and the dynamics of preferential flow paths in porous media. *Biofouling*, 29(9):1069–1086, 2013.

A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.

S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

J. Brynjarsdóttir and A. O'Hagan. Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30(11):114007, 2014.

C. J. Burges. *Geometric Methods for Feature Extraction and Dimensional Reduction - A Guided Tour*, pages 53–82. Springer US, Boston, MA, 2010. ISBN 978-0-387-09823-4. doi: 10.1007/978-0-387-09823-4_4. URL https://doi.org/10.1007/978-0-387-09823-4_4.

K. P. Burnham and D. R. Anderson. Practical use of the information-theoretic approach. In *Model Selection and Inference*, pages 75–117. Springer, 1998.

K. P. Burnham and D. R. Anderson. A Practical Information-Theoretic Approach. *Model Selection and Multimodel Inference, Springer (second edition), New York*, 2, 2002.

K. P. Burnham and D. R. Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

T. Carraro, C. Goll, A. Marciniak-Czochra, and A. Mikelić. Effective interface conditions for the forced infiltration of a viscous fluid into a porous medium using homogenization. *Computer Methods in Applied Mechanics and Engineering*, 292:195–220, 2015.

M.-H. Chen, Q.-M. Shao, and J. G. Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media, 2012.

W. Chen, L. Baghdasaryan, T. Buranathiti, and J. Cao. Model validation via uncertainty propagation and data transformations. *AIAA journal*, 42(7):1406–1415, 2004.

W. Chen, Y. Xiong, K.-L. Tsui, and S. Wang. A design-driven validation approach using bayesian prediction models. *Journal of Mechanical Design*, 130(2), 12 2007.

K. Cheng and Z. Lu. Structural reliability analysis based on ensemble learning of surrogate models. *Structural Safety*, 83:101905, 2020.

G. Claeskens. Statistical model choice. *Annual review of statistics and its application*, 3:233–256, 2016.

H. W. Coleman and F. Stern. Uncertainties and CFD code validation. *Journal of Fluids Engineering*, 119(4):795–803, 12 1997. ISSN 0098-2202. doi: 10.1115/1.2819500.

G. Consonni, D. Fouskakis, B. Liseo, and I. Ntzoufras. Prior distributions for objective bayesian analysis. *Bayesian Analysis*, 13(2):627–679, 2018.

P. G. Constantine. *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM, 2015.

M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

K. Crombecq. *Surrogate modeling of computer experiments with sequential experimental design*. PhD thesis, Ghent University, 2011.

A. B. Cunningham, H. Class, A. Ebigbo, R. Gerlach, A. J. Phillips, and J. Hommel. Field-scale modeling of microbially induced calcite precipitation. *Computational Geosciences*, 23(2):399–414, 2019.

M. O. Cuthbert, L. A. McMillan, S. Handley-Sidhu, M. S. Riley, D. J. Tobler, and V. R. Phoenix. A field and modeling study of fractured rock permeability reduction using microbially induced calcite precipitation. *Environmental Science & Technology*, 47 (23):13637–13643, 2013.

I. J. Dahabreh, J. A. Chan, A. Earley, D. Moorthy, E. E. Avendano, T. A. Trikalinos, E. M. Balk, and J. B. Wong. A review of validation and calibration methods for health care modeling and simulation. *Modeling and Simulation in the Context of Health Technology Assessment: Review of Existing Guidance, Future Research Needs, and Validity Assessment*, 2017.

B. Debusschere. Intrusive polynomial chaos methods for forward uncertainty propagation. Technical report, Sandia National Lab.(SNL-CA), Livermore, CA (United States), 2015.

D. Den Hertog, J. P. Kleijnen, and A. Y. Siem. The correct Kriging variance estimated by bootstrapping. *Journal of the Operational Research Society*, 57(4):400–409, 2006.

D. Deschrijver, K. Crombecq, H. M. Nguyen, and T. Dhaene. Adaptive sampling algorithm for macromodeling of parameterized parameter responses. *IEEE Transactions on Microwave Theory and Techniques*, 59(1):39–45, 2011.

M. Discacciati and L. Gerardo-Giorda. Optimized Schwarz methods for the Stokes–Darcy coupling. *IMA Journal of Numerical Analysis*, 38:1959–1983, 2018.

M. Discacciati and A. Quarteroni. Navier–Stokes/Darcy coupling: modeling, analysis, and numerical approximation. *Revista Matemática Complutense*, 22:315–426, 2009.

M. Discacciati, E. Miglio, and A. Quarteroni. Mathematical and numerical models for coupling surface and groundwater flows. *Applied Numerical Mathematics*, 43:57–74, 2002.

A. Doostan and H. Owhadi. A non-adapted sparse approximation of pdes with stochastic inputs. *Journal of Computational Physics*, 230(8):3015–3034, 2011.

K. Dowding. Quantitative validation of mathematical models. In *ASME International Mechanical Engineering Congress and Exposition*, volume 35623, pages 237–250. American Society of Mechanical Engineers, 2001.

K. J. Dowding and B. M. Rutherford. An approach to model validation and model-based prediction–polyurethane foam case study. Technical report, Sandia National Laboratories, Albuquerque, NM, and Livermore, CA, 2003.

K. J. Dowding, I. H. Leslie, M. L. Hobbs, B. M. Rutherford, R. G. Hills, and M. M. Pilch. Case study for model validation: assessing a model for thermal decomposition of polyurethane foam. Technical report, Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA, 2004.

D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):45–70, 1995.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

S. Dubreuil, M. Berveiller, F. Petitjean, and M. Salaün. Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion. *Reliability Engineering & System Safety*, 121:263–275, 2014.

S. Dupraz, M. Parmentier, B. Ménez, and F. Guyot. Experimental and numerical modeling of bacterially induced pH increase and calcite precipitation in saline aquifers. *Chemical Geology*, 265(1-2):44–53, July 2009. ISSN 00092541. doi: 10.1016/j.chemgeo.2009.05.003.

A. Ebigbo, A. J. Phillips, R. Gerlach, R. Helmig, A. B. Cunningham, H. Class, and L. H. Spangler. Darcy-scale modeling of microbially induced carbonate mineral precipitation in sand columns. *Water Resources Research*, 48(7):W07519, July 2012. ISSN 0043-1397. doi: 10.1029/2011WR011714.

W. Edwards, H. Lindman, and L. J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193, 1963.

B. Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

E. Eggenweiler and I. Rybak. Unsuitability of the Beavers–Joseph interface condition for filtration problems. *Journal of Fluid Mechanics*, 892:A10, 2020.

E. Eggenweiler and I. Rybak. Effective coupling conditions for arbitrary flows in Stokes–Darcy systems. *Multiscale Modeling & Simulation*, 19(2):731–757, 2021.

L. P. Elliott and B. W. Brook. Revisiting Chamberlin: multiple working hypotheses for the 21st century. *BioScience*, 57(7):608–614, 2007.

A. H. Elsheikh, I. Hoteit, and M. F. Wheeler. Efficient Bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates. *Computer Methods in Applied Mechanics and Engineering*, 269:515–537, 2014.

T. Enemark, L. J. Peeters, D. Mallants, and O. Batelaan. Hydrogeological conceptual model building and testing: A review. *Journal of Hydrology*, 569:310–329, 2019.

N. Fajraoui, S. Marelli, and B. Sudret. Sequential design of experiment for sparse polynomial chaos expansions. *SIAM/ASA Journal on Uncertainty Quantification*, 5 (1):1061–1085, 2017.

A. C. Faul and M. E. Tipping. Analysis of sparse Bayesian learning. In *Advances in neural information processing systems*, pages 383–389, 2002.

J. Faust, S. Gilchrist, J. H. Wright, and E. Zakrajšsek. Credit spreads as predictors of real-time economic activity: a Bayesian model-averaging approach. *Review of Economics and Statistics*, 95(5):1501–1519, 2013.

S. Ferson, V. Kreinovick, L. Ginzburg, and F. Sentz. Constructing probability boxes and dempster-shafer structures. Technical report, Sandia National Lab, Albuquerque, NM (US), 2003.

S. Ferson, W. L. Oberkampf, and L. Ginzburg. Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*, 197(29-32):2408–2430, 2008.

D. Foreman-Mackey, W. M. Farr, M. Sinha, A. M. Archibald, D. W. Hogg, J. S. Sanders, J. Zuntz, P. K. Williams, A. R. Nelson, M. de Val-Borro, et al. emcee

v3: A Python ensemble sampling toolkit for affine-invariant MCMC. *arXiv preprint arXiv:1911.07688*, 2019.

N. Friel and J. Wyse. Estimating the evidence–a review. *Statistica Neerlandica*, 66(3): 288–308, 2012.

P. Gardner, T. Rogers, C. Lord, and R. Barthorpe. Learning model discrepancy: A Gaussian process and sampling-based approach. *Mechanical Systems and Signal Processing*, 152:107381, 2021.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

S. Geisser and W. F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

A. E. Gelfand and D. K. Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3): 501–514, 1994.

A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

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

J. Geweke. Bayesian model comparison and validation. *American Economic Review*, 97(2):60–64, 2007.

M. Ghaith, Z. Li, and B. W. Baetz. Uncertainty analysis for hydrological models with interdependent parameters: An improved polynomial chaos expansion approach. *Water Resources Research*, 57(8):e2020WR029149, 2021.

R. G. Ghanem and P. D. Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.

W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

A. Giunta, S. Wojtkiewicz, and M. Eldred. Overview of modern design of experiments methods for computational simulations. In *41st Aerospace Sciences Meeting and Exhibit*, page 649, 2003.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

M. G. Gomez, C. M. Anderson, C. M. R. Graddy, J. T. DeJong, D. C. Nelson, and T. R. Ginn. Large-scale comparison of bioaugmentation and biostimulation approaches for biocementation of sands. *Journal of Geotechnical and Geoenvironmental Engineering*, 143(5):04016124, 2017. doi: 10.1061/(ASCE)GT.1943-5606.0001640.

J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5(1):65–80, 2010.

B. Goyeau, D. Lhuillier, D. Gobin, and M. Velarde. Momentum transport at a fluid-porous interface. *International Journal of Heat and Mass Transfer*, 46:4071–4081, 2003.

Q. F. Gronau and E.-J. Wagenmakers. Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior*, 2(1):1–11, 2019a.

Q. F. Gronau and E.-J. Wagenmakers. Rejoinder: More limitations of Bayesian leave-one-out cross-validation. *Computational Brain & Behavior*, 2(1):35–47, 2019b.

S. F. Gull. Developments in maximum entropy data analysis. In *Maximum entropy and Bayesian methods*, pages 53–71. Springer, 1989.

X. Guo, Q. Sun, D. Dias, and E. Antoinet. Probabilistic assessment of an earth dam stability design using the adaptive polynomial chaos expansion. *Bulletin of Engineering Geology and the Environment*, 79(9):4639–4655, 2020.

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), 2012. doi: 10.1029/2011WR011044.

H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

W. Hackbusch. On first and second order box schemes. *Computing*, 41(4):277–296, 1989.

M. Hadigol and A. Doostan. Least squares polynomial chaos expansion: A review of sampling strategies. *Computer Methods in Applied Mechanics and Engineering*, 332: 382–407, 2018.

A. Hald. *A History of Mathematical Statistics from 1750 to 1930*, volume 2. Wiley New York, 1998.

J. H. Halton. Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702, 1964.

N. Hamdan, E. Kavazanjian Jr, and B. E. Rittmann. Sequestration of radionuclides and metal contaminants through microbially-induced carbonate precipitation. In *Proc. 14th Pan American Conf. Soil Mech. Geotech. Engng, Toronto*, 2011.

J. Hammersley. *Monte Carlo methods*. Springer Science & Business Media, 2013.

J. Hampton and A. Doostan. Compressive sampling methods for sparse polynomial chaos expansions. *Handbook of Uncertainty Quantification*, pages 1–29, 2016.

C. Han and B. P. Carlin. MCMC methods for computing Bayes factors: A comparative review. *Biometrika*, 82(4):711–732, 2000.

N. Hanspal, A. Waghode, V. Nassehi, and R. Wakeman. Development of a predictive mathematical model for coupled Stokes/ Darcy flows in cross-flow membrane filtration. *Chemical Engineering Journal*, 149:132–142, 2009.

F. H. Harlow and J. E. Welch. Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *The Physics of Fluids*, 8(12):2182–2189, 1965.

W. V. Harper and S. K. Gupta. *Sensitivity/uncertainty analysis of a borehole scenario comparing Latin hypercube sampling and deterministic sensitivity approaches*. Office of Nuclear Waste Isolation, Battelle Memorial Institute Columbus, Ohio, 1983.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97.

I. M. Head. Bioremediation: towards a credible technology. *Microbiology*, 144(3): 599–608, 1998.

R. Helmig. *Multiphase Flow and Transport Processes in the Subsurface - A Contribution to the Modeling of Hydrosystems*. Springer Verlag, 1997.

D. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103 (482):570–583, 2008.

R. G. Hills. Model validation: model parameter and measurement uncertainty. *Journal of Heat Transfer*, 128(4):339–351, 10 2005. ISSN 0022-1481. doi: 10.1115/1.2164849. URL `https://doi.org/10.1115/1.2164849`.

R. G. Hills and I. H. Leslie. Statistical validation of engineering and scientific models: validation experiments to application. Technical report, Sandia National Lab, Albuquerque, NM (US), 2003.

R. G. Hills and T. G. Trucano. Statistical validation of engineering and scientific models: Background. *Sandia National Laboratories, SAND99-1256*, 36, 1999.

R. G. Hills and T. G. Trucano. Statistical validation of engineering and scientific models: a maximum likelihood based metric. Technical report, Sandia National Lab., Albuquerque, NM (US), 2002.

M. Hinne, Q. F. Gronau, D. van den Bergh, and E.-J. Wagenmakers. A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2):200–215, 2020.

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.

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, 2016.

M. B. Hooten and N. T. Hobbs. A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28, 2015.

U. Hornung. *Homogenization and Porous Media*. Springer, New York, 1996.

K. T. Hu and G. E. Orient. The 2014 Sandia verification and validation challenge: Problem statement. *Journal of Verification, Validation and Uncertainty Quantification*, 1(1), 2016.

S. Huang, M. Cao, and L. Cheng. Experimental study on the mechanism of enhanced oil recovery by multi-thermal fluid in offshore heavy oil. *International Journal of Heat and Mass Transfer*, 122:1074–1084, 2018.

K. S. Hunter, Y. Wang, and P. Van Cappellen. Kinetic modeling of microbially-driven redox chemistry of subsurface environments: coupling transport, microbial metabolism and geochemistry. *Journal of Hydrology*, 209(1-4):53–80, 1998.

M. Hutter. On universal prediction and bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007.

J. T. Hwang, G. Casella, C. Robert, M. T. Wells, and R. H. Farrell. Estimation of accuracy in testing. *The annals of statistics*, pages 490–509, 1992.

A. Højberg and J. Refsgaard. Model uncertainty – parameter uncertainty versus conceptual models. *Water Science and Technology*, 52(6):177–186, Sept. 2005. ISSN 0273-1223, 1996-9732. doi: 10.2166/wst.2005.0166.

B. Iooss and P. Lemaître. A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pages 101–122. Springer, 2015.

W. Jäger and A. Mikelić. On the interface boundary conditions by Beavers, Joseph and Saffman. *SIAM Journal on Applied Mathematics*, 60:1111–1127, 2000.

W. Jäger and A. Mikelić. Modeling effective interface laws for transport phenomena between an unconfined fluid and a porous medium using homogenization. *Transport in Porous Media*, 78:489–508, 2009.

A. Jarauta, V. Zingan, P. Minev, and M. Secanell. A compressible fluid flow model coupling channel and porous media flows and its application to fuel cell materials. *Transport in Porous Media*, 134(2):351–386, 2020.

E. T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

H. Jeffreys. *The theory of probability*. Oxford Univ. Press, Oxford, 1961.

B. C. Jeon, J. H. Jung, B. D. Youn, Y.-W. Kim, and Y.-C. Bae. Datum unit optimization for robustness of a journal bearing diagnosis system. *International Journal of Precision Engineering and Manufacturing*, 16(11):2411–2425, 2015.

S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on signal processing*, 56(6):2346–2356, 2008.

X. Jiang and S. Mahadevan. Bayesian risk-based decision method for model validation under uncertainty. *Reliability Engineering & System Safety*, 92(6):707–718, 2007.

X. Jiang and S. Mahadevan. Bayesian validation assessment of multivariate computational models. *Journal of Applied Statistics*, 35(1):49–65, 2008a.

X. Jiang and S. Mahadevan. Bayesian wavelet method for multivariate model assessment of dynamic systems. *Journal of Sound and Vibration*, 312(4-5):694–712, 2008b.

X. Jiang and S. Mahadevan. Bayesian inference method for model validation and confidence extrapolation. *Journal of Applied Statistics*, 36(6):659–677, 2009a.

X. Jiang and S. Mahadevan. Bayesian structural equation modeling method for hierarchical model validation. *Reliability Engineering & System Safety*, 94(4):796–809, 2009b.

X. Jiang, X. Cheng, and Y. Yuan. *Validation Using Bayesian Methods*, pages 497–524. Springer International Publishing, Cham, 2019. ISBN 978-3-319-70766-2. doi: 10.1007/978-3-319-70766-2_20. URL `https://doi.org/10.1007/978-3-319-70766-2_20`.

X. Jiang, R. Ma, Y. Wang, W. Gu, W. Lu, and J. Na. Two-stage surrogate model-assisted bayesian framework for groundwater contaminant source identification. *Journal of Hydrology*, 594:125955, 2021.

R. A. Johnson, I. Miller, and J. E. Freund. *Probability and statistics for engineers*, volume 2000. Pearson Education London, 2000.

D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

C.-J. Kat and P. S. Els. Validation metric based on relative error. *Mathematical and Computer Modelling of Dynamical Systems*, 18(5):487–520, 2012.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

C. Keysers, V. Gazzola, and E.-J. Wagenmakers. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, 23(7):788–799, 2020.

B. Kim, Y. Lee, and D.-H. Choi. Construction of the radial basis function based on a sequential sampling approach using cross-validation. *Journal of Mechanical Science and Technology*, 23(12):3357–3365, 2009.

T. Kim and B. D. Youn. Identifiability-based model decomposition for hierarchical calibration. *Structural and Multidisciplinary Optimization*, 60(5):1801–1811, 2019.

C. M. Kirkland, A. Thane, R. Hiebert, R. Hyatt, J. Kirksey, A. B. Cunningham, R. Gerlach, L. Spangler, and A. J. Phillips. Addressing wellbore integrity and thief zone permeability using microbially-induced calcium carbonate precipitation (MICP): A field demonstration. *Journal of Petroleum Science and Engineering*, 190:107060, 2020. ISSN 0920-4105. doi: 10.1016/j.petrol.2020.107060.

J. P. Kleijnen and W. C. Van Beers. Monotonicity-preserving bootstrapped Kriging metamodels for expensive simulations. *Journal of the Operational Research Society*, 64(5):708–717, 2013.

K.-R. Koch. *Parameter estimation and hypothesis testing in linear models*. Springer Science & Business Media, 1999.

T. Koch, D. Gläser, K. Weishaupt, S. Ackermann, M. Beck, B. Becker, S. Burbulla, H. Class, E. Coltman, S. Emmert, T. Fetzer, C. Grüninger, K. Heck, J. Hommel, T. Kurz, M. Lipp, F. Mohammadi, S. Scherrer, M. Schneider, G. Seitz, L. Stadler, M. Utz, F. Weinhardt, and B. Flemisch. Dumux 3–an open-source simulator for solving flow and transport problems in porous media with a focus on model coupling. *Computers & Mathematics with Applications*, 81:423–443, 2021a.

T. Koch, K. Weishaupt, J. Müller, B. Weigand, and R. Helmig. A (dual) network model for heat transfer in porous media. *Transport in Porous Media*, 140(1):107–141, Oct 2021b. ISSN 1573-1634.

A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

C. Kuhn and R. Müller. A continuum phase field model for fracture. *Engineering Fracture Mechanics*, 77(18):3625–3634, 2010.

S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

A. Kupresanin and G. Johannesson. Comparison of sequential designs of computer experiments in high dimensions. Technical report, Lawrence Livermore National Lab, Livermore, CA (US), 2011.

U. Lācis and S. Bagheri. A framework for computing effective boundary conditions at the interface between free fluid and a porous medium. *Journal of Fluid Mechanics*, 812:866–889, 2017.

U. Lācis, Y. Sudhakar, S. Pasche, and S. Bagheri. Transfer of mass and momentum at rough and porous surfaces. *J. Fluid Mech.*, 884:A21, 2020.

C. Q. Lam. *Sequential adaptive designs in computer experiments for response surface model fit*. PhD thesis, The Ohio State University, 2008.

D. Landa-Marbán, S. Tveit, K. Kumar, and S. E. Gasda. Practical approaches to study microbially induced calcite precipitation at the field scale. *International Journal of Greenhouse Gas Control*, 106:103256, 2021.

J. Law. Bayesian statistics and functional programming: Bayesian inference using rejection sampling, 2019. URL `https://jonnylaw.rocks/posts/2019-02-25-rejection_sampling/`.

T. Le and B. Clarke. A Bayes interpretation of stacking for m-complete and m-open settings. *Bayesian Analysis*, 12(3):807–829, 2017.

M. Le Bars and M. Worster. Interfacial conditions between a pure fluid and a porous medium: implications for binary alloy solidification. *Journal of Fluid Mechanics*, 550:149–173, 2006.

E. E. Leamer. *Specification searches: Ad hoc inference with non–experimental data*, volume 53. John Wiley & Sons Incorporated, 1978.

G. Lee, W. Kim, H. Oh, B. D. Youn, and N. H. Kim. Review of statistical model calibration and validation—from the perspective of uncertainty structures. *Structural and Multidisciplinary Optimization*, 60(4):1619–1644, 2019.

J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*, volume 1. Springer, 2007.

E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

Y. Ling and S. Mahadevan. Quantitative model validation techniques: New insights. *Reliability Engineering & System Safety*, 111:217–231, 2013.

Y. Ling, J. Mullins, and S. Mahadevan. Selection of model discrepancy priors in Bayesian calibration. *Journal of Computational Physics*, 276:665–680, 2014.

H. Liu, S. Xu, Y. Ma, X. Chen, and X. Wang. An adaptive Bayesian sequential sampling approach for global metamodeling. *Journal of Mechanical Design*, 138(1), 2016a.

H. Liu, S. Xu, X. Wang, J. Meng, and S. Yang. Optimal weighted pointwise ensemble of radial basis functions with different basis functions. *AIAA Journal*, pages 3117–3133, 2016b.

H. Liu, Y.-S. Ong, and J. Cai. A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design. *Structural and Multidisciplinary Optimization*, 57(1):393–416, 2018.

J. S. Liu and J. S. Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

Y. Liu, W. Chen, P. Arendt, and H.-Z. Huang. Toward a better understanding of model validation metrics. *Journal of Mechanical Design*, 133(7), 2011.

D. R. Lovley and F. H. Chapelle. Deep subsurface microbial processes. *Reviews of Geophysics*, 33(3):365–381, 1995.

N. Lüthen, S. Marelli, and B. Sudret. Sparse polynomial chaos expansions: Literature survey and benchmark. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2): 593–649, 2021.

D. J. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992a.

D. J. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992b.

K. T. B. MacQuarrie and K. U. Mayer. Reactive transport modeling in fractured rock: A state-of-the-science review. *Earth-Science Reviews*, 72(3-4):189–227, 2005. ISSN 00128252. doi: 10.1016/j.earscirev.2005.07.003.

S. Mahadevan and R. Rebba. Validation of reliability computational models using bayes networks. *Reliability Engineering & System Safety*, 87(2):223–232, 2005.

C. V. Mai and B. Sudret. Surrogate models for oscillatory systems using sparse polynomial chaos expansions and stochastic time warping. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):540–571, 2017.

D. Maljovec, B. Wang, A. Kupresanin, G. Johannesson, V. Pascucci, and P.-T. Bremer. Adaptive sampling with topological scores. *International Journal for Uncertainty Quantification*, 3(2), 2013.

P. Manfredi and R. Trinchero. A data compression strategy for the efficient uncertainty quantification of time-domain circuit responses. *IEEE Access*, 8:92019–92027, 2020.

S. Marelli and B. Sudret. An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis. *Structural Safety*, 75:67–74, 2018.

S. Marelli, N. Lüthen, and B. Sudret. UQLab user manual – Polynomial chaos expansions. Technical report, Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland, 2021. Report # UQLab-V1.4-104.

J.-M. Marin and C. P. Robert. Importance sampling methods for Bayesian discrimination between embedded models. *arXiv preprint arXiv:0910.2325*, 2009.

K. A. Maupin, L. P. Swiler, and N. W. Porter. Validation metrics for deterministic and probabilistic data. *Journal of Verification, Validation and Uncertainty Quantification*, 3(3), 2018.

M. J. McInerney, D. P. Nagle, and R. M. Knapp. Microbially Enhanced Oil Recovery: Past, Present, and Future. *Petroleum Microbiology*, pages 215–237, 2005.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.

M. Megharaj, B. Ramakrishnan, K. Venkateswarlu, N. Sethunathan, and R. Naidu. Bioremediation approaches for organic pollutants: A critical perspective. *Environment International*, 37(8):1362–1375, 2011.

E. Mehdad and J. P. Kleijnen. Classic Kriging versus Kriging with bootstrapping or conditional simulation: classic Kriging's robust confidence intervals and optimization. *Journal of the Operational Research Society*, 66(11):1804–1814, 2015.

Y. Mehmani and M. T. Balhoff. Bridging from pore to continuum: a hybrid mortar domain decomposition framework for subsurface flow and transport. *Multiscale Modeling & Simulation*, 12(2):667–693, 2014.

Y. Mehmani and H. A. Tchelepi. Minimum requirements for predictive pore-network modeling of solute transport in micromodels. *Advances in Water Research.*, 108: 83–98, 2017.

Z. A. Memon, R. Trinchero, P. Manfredi, F. Canavero, and I. S. Stievano. Compressed machine learning models for the uncertainty quantification of power distribution networks. *Energies*, 13(18):4881, 2020.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

M. Mierzwiczak, A. Fraska, and J. Grabski. Determination of the slip constant in the Beavers–Joseph experiment for laminar fluid flow through porous media using a meshless method. *Mathematical Problems in Engineering*, 2019:1494215, 2019.

J. M. Minto, R. J. Lunn, and G. El Mountassir. Development of a Reactive Transport Model for Field-Scale Simulation of Microbially Induced Carbonate Precipitation. *Water Resources Research*, 55(8):7229–7245, 2019. doi: 10.1029/2019WR025153.

A. C. Mitchell, A. J. Phillips, L. Schultz, S. Parks, L. H. Spangler, A. B. Cunningham, and R. Gerlach. 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, July 2013. ISSN 17505836. doi: 10.1016/j.ijggc.2013.02.001.

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, 2018.

F. Mohammadi, E. Eggenweiler, B. Flemisch, S. Oladyshkin, I. Rybak, M. Schneider, and K. Weishaupt. A surrogate-assisted uncertainty-aware Bayesian validation framework and its application to coupling free flow and porous-medium flow. Submitted to *Computational Geosciences*, 2022. URL `https://arxiv.org/abs/2106.13639`.

R. D. Moser and T. A. Oliver. Validation of physical models in the presence of uncertainty. *Handbook of Uncertainty Quantification*, pages 1–28, 2015.

C. N. Mulligan and R. Galvez-Cloutier. Bioremediation of metal contamination. *Environmental Monitoring and Assessment*, 84(1-2):45–60, 2003.

J. B. Nagel and B. Sudret. Hamiltonian Monte Carlo and borrowing strength in hierarchical inverse problems. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 2(3):B4015008, 2016.

J. B. Nagel, J. Rieckermann, and B. Sudret. Principal component analysis and sparse polynomial chaos expansions for global sensitivity analysis and model calibration: Application to urban drainage simulation. *Reliability Engineering & System Safety*, 195:106737, 2020.

M. K. Nassar, D. Gurung, M. Bastani, T. R. Ginn, B. Shafei, M. G. Gomez, C. M. Graddy, D. C. Nelson, and J. T. DeJong. Large-scale experiments in microbially induced calcite precipitation (MICP): Reactive transport model development and prediction. *Water Resources Research*, 54(1):480–500, 2018.

T. H. Naylor and J. M. Finger. Verification of computer simulation models. *Management Science*, 14(2):B–92, 1967.

R. M. Neal et al. MCMC using hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2(11):2, 2011.

S. P. Neuman. Maximum likelihood bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment*, 17(5):291–305, 2003.

S. P. Neuman. Trends, prospects and challenges in quantifying flow and transport through fractured rocks. *Hydrogeology Journal*, 13(1):124–147, 2005.

E. Nikolaidis and R. T. Haftka. Theories of uncertainty for risk assessment when data is scarce. *International Journal for Advanced Manufacturing Systems*, 4(1):49–56, 2001.

A. Notin, N. Gayton, J. L. Dulong, M. Lemaire, P. Villon, and H. Jaffal. Rpcm: a strategy to perform reliability analysis using polynomial chaos and resampling: Application to fatigue design. *European Journal of Computational Mechanics/Revue Européenne de Mécanique Numérique*, 19(8):795–830, 2010.

J. E. Oakley and A. O'Hagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769, 2004.

W. Oberkampf and T. Trucano. Validation methodology in computational fluid dynamics. In *Fluids 2000 Conference and Exhibit*, page 2549, 2000.

W. Oberkampf, J. Helton, and K. Sentz. Mathematical representation of uncertainty. In *19th AIAA applied aerodynamics conference*, page 1645, 2001.

W. L. Oberkampf and M. F. Barone. Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics*, 217(1):5–36, 2006.

W. L. Oberkampf and C. J. Roy. *Verification and validation in scientific computing*. Cambridge University Press, 2010.

W. L. Oberkampf and T. G. Trucano. Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences*, 38(3):209–272, 2002.

W. L. Oberkampf and T. G. Trucano. Verification and validation benchmarks. *Nuclear Engineering and Design*, 238(3):716–743, 2008.

W. L. Oberkampf, T. G. Trucano, and C. Hirsch. Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*, 57 (5):345–384, 2004.

A. J. Ochoa-Tapia and S. Whitaker. Momentum transfer at the boundary between a porous medium and a homogeneous fluid. I: theoretical development. *International Journal of Heat Mass Transfer*, 38:2635–2646, 1995.

H. Oh, H. Choi, J. H. Jung, and B. D. Youn. A robust and convex metric for unconstrained optimization in statistical model calibration—probability residual (PR). *Structural and Multidisciplinary Optimization*, 60(3):1171–1187, 2019.

S. Oladyshkin and W. Nowak. Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. *Reliability Engineering & System Safety*, 106: 179–190, 2012.

S. Oladyshkin and W. Nowak. The connection between Bayesian inference and information theory for model selection, information gain and experimental design. *Entropy*, 21(11):1081, 2019.

S. Oladyshkin, F. De 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.

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.

S. Oladyshkin, F. Mohammadi, I. Kroeker, and W. Nowak. Bayesian3 active learning for the gaussian process emulator using information theory. *Entropy*, 22(8):890, 2020.

M. Oostrom, Y. Mehmani, P. Romero-Gomez, Y. Tang, H. Liu, H. Yoon, Q. Kang, V. Joekar-Niasar, M. Balhoff, T. Dewers, and C. Zhang. Pore-scale and continuum simulations of solute transport micromodel benchmark experiments. *Computational Geosciences*, 20(4):857–879, 2016.

T. L. Paez and A. Urbina. Validation of mathematical models of complex structural dynamic systems. In *Proceedings of the ninth international congress on sound and vibration, Orlando, FL*, 2002.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.

T. W. Patzek and D. B. Silin. Shape factor and hydraulic conductance in noncircular capillaries: I. One-phase creeping flow. *Journal of Colloid and Interface Science*, 236: 295–304, 2001.

F. Pérez-Cruz. Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory*, pages 1666–1670. IEEE, 2008.

A. J. Phillips, E. Lauchnor, J. Eldring, R. Esposito, A. C. Mitchell, R. Gerlach, A. B. Cunningham, and L. H. Spangler. Potential CO_2 leakage reduction through biofilm-induced calcium carbonate precipitation. *Environmental Science & Technology*, 47 (1):142–149, 2013.

A. J. Phillips, A. B. Cunningham, R. Gerlach, R. Hiebert, C. Hwang, B. P. Lomans, J. Westrich, C. Mantilla, J. Kirksey, R. Esposito, and L. H. Spangler. Fracture sealing with microbially-induced calcium carbonate precipitation: A field study. *Environmental Science & Technology*, 50:4111–4117, 2016. doi: 10.1021/acs.est.5b05559.

V. Picheny, N. H. Kim, and R. T. Haftka. Application of bootstrap method in conservative estimation of reliability with limited samples. *Structural and Multidisciplinary Optimization*, 41(2):205–217, 2010.

S. J. Prince. *Computer vision: models, learning, and inference*. Cambridge University Press, 2012.

N. V. Queipo, R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. K. Tucker. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, 41(1): 1–28, 2005.

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

A. E. Raftery and J. D. Banfield. Stopping the gibbs sampler, the use of morphology, and other issues in spatial statistics (Bayesian image restoration, with two applications in spatial statistics)–(Discussion). *Annals of the Institute of Statistical Mathematics*, 43(1):p32–43, 1991.

A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.

S. Ranftl, W. von der Linden, and M. . S. Committee. Bayesian surrogate analysis and uncertainty propagation. In *Physical Sciences Forum*, volume 3, page 6. MDPI, 2021.

R. Rebba and S. Mahadevan. Computational methods for model reliability assessment. *Reliability Engineering & System Safety*, 93(8):1197–1207, 2008.

R. Rebba, S. Huang, Y. Liu, and S. Mahadevan. Statistical validation of simulation models. *International Journal of Materials and Product Technology*, 25(1-3):164–181, 2006.

J. C. Refsgaard and H. J. Henriksen. Modelling guidelines–terminology and guiding principles. *Advances in Water Resources*, 27(1):71–82, 2004.

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, 2012.

B. Renard, D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, 46(5), 2010. doi: 10.1029/2009WR008328.

M. L. Rizzo and G. J. Székely. Energy distance. *Wiley Interdisciplinary Reviews: Computational statistics*, 8(1):27–38, 2016.

P. J. Roache. *Verification and validation in computational science and engineering*, volume 895. Hermosa Albuquerque, NM, 1998.

C. P. Robert and D. Wraith. Computational methods for Bayesian model choice. In *AIP conference proceedings*, volume 1193, pages 251–262. American Institute of Physics, 2009.

C. P. Robert, G. Casella, and G. Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

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), 2008. ISSN 00431397. doi: 10.1029/2008WR006908.

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

S. M. Ross. *Introduction to probability and statistics for engineers and scientists*. Academic press, 2020.

I. Rybak, C. Schwarzmeier, E. Eggenweiler, and U. Rüde. Validation and calibration of coupled porous-medium and free-flow problems using pore-scale resolved models. *Computational Geosciences*, 25(2):621–635, 2021. ISSN 1420-0597.

J. Sacks, S. B. Schiller, and W. J. Welch. Designs for computer experiments. *Technometrics*, 31(1):41–47, 1989.

S. Sankararaman and S. Mahadevan. Model validation under epistemic uncertainty. *Reliability Engineering & System Safety*, 96(9):1232–1241, 2011.

S. Sankararaman and S. Mahadevan. Assessing the reliability of computational models under uncertainty. In *54th AIAA/ASME/ASCE/AHS/ASC Structures, structural dynamics, and materials conference*, page 1873, 2013.

S. Sankararaman and S. Mahadevan. Integration of model verification, validation, and calibration for uncertainty quantification in engineering systems. *Reliability Engineering & System Safety*, 138:194–209, 2015.

S. Sankararaman, Y. Ling, C. Shantz, and S. Mahadevan. Inference of equivalent initial flaw size under multiple sources of uncertainty. *International Journal of Fatigue*, 33 (2):75–89, 2011.

L. K. Saul, K. Q. Weinberger, F. Sha, J. Ham, and D. D. Lee. Spectral methods for dimensionality reduction. *Semi-Supervised Learning*, 3, 2006.

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):e2020WR028100, 2020.

A. Schäfer Rodrigues Silva, T. K. 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*, pages 1–33, 2022.

T. D. Scheibe, E. M. Murphy, X. Chen, A. K. Rice, K. C. Carroll, B. J. Palmer, A. M. Tartakovsky, I. Battiato, and B. D. Wood. An analysis platform for multiscale hydrogeologic modeling with emphasis on hybrid multiscale methods. *Groundwater*, 53(1):38–56, 2015.

S. Scheurer, A. Schäfer Rodrigues Silva, F. Mohammadi, J. Hommel, S. Oladyshkin, B. Flemisch, and W. Nowak. Surrogate-based Bayesian comparison of computationally expensive models: application to microbially induced calcite precipitation. *Computational Geosciences*, 25(6):1899–1917, 2021.

A. Schöniger. *Bayesian assessment of conceptual uncertainty in hydrosystem modeling*. PhD thesis, Universitätsbibliothek Tübingen, 2016.

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, 2014.

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, 2015a.

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

G. Schoups and J. A. Vrugt. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-gaussian errors. *Water Resources Research*, 46(10), 2010.

D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

M. W. Seeger and H. Nickisch. Compressed sensing and Bayesian experimental design. In *Proceedings of the 25th international conference on Machine learning*, pages 912–919, 2008.

B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

T. W. Simpson, D. K. Lin, and W. Chen. Sampling strategies for computer experiments: design and analysis. *International Journal of Reliability and Applications*, 2(3):209–240, 2001.

T. W. Simpson, A. J. Booker, D. Ghosh, A. A. Giunta, P. N. Koch, and R.-J. Yang. Approximation methods in multidisciplinary analysis and optimization: a panel discussion. *Structural and multidisciplinary optimization*, 27(5):302–313, 2004.

P. Singh, D. Deschrijver, and T. Dhaene. A balanced sequential design strategy for global surrogate modeling. In *2013 Winter Simulations Conference (WSC)*, pages 2172–2179. IEEE, 2013.

M. Sinsbeck, E. Cooke, and W. Nowak. Sequential design of computer experiments for the computation of Bayesian model evidence. *SIAM/ASA Journal on Uncertainty Quantification*, 9(1):260–279, 2021.

A. Smith, P. A. Naik, and C.-L. Tsai. Markov-switching model selection using Kullback–Leibler divergence. *Journal of Econometrics*, 134(2):553–577, 2006.

I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4):784–802, 1967.

I. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling Computational Experiments*, 1(4):407–414, 1993.

A. Sokal. Monte Carlo methods in statistical mechanics: Foundations and new algorithms note to the reader. In *Functional integration*, pages 131–192. Springer, 1997.

H. Son, G. Lee, K. Kang, Y.-J. Kang, B. D. Youn, I. Lee, and Y. Noh. Industrial issues and solutions to statistical model improvement: a case study of an automobile steering column. *Structural and Multidisciplinary Optimization*, 61(4):1739–1756, 2020.

C. Steefel and K. MacQuarrie. *Reactive Transport in Porous Media. Reviews in Mineralogy*, chapter Approaches to modelling of reactive transport in porous media, page 82–129. Mineralogical Society of America, Washington, 1996.

C. Steefel, D. Depaolo, and P. Lichtner. Reactive transport modeling: An essential tool and a new research approach for the Earth sciences. *Earth and Planetary Science Letters*, 240(3-4):539–558, 2005. ISSN 0012821X. doi: 10.1016/j.epsl.2005.09.017.

G. W. Stewart. *Afternotes on numerical analysis*. SIAM, 1996.

S. Stocks-Fischer, J. K. Galinat, and S. S. Bang. Microbiological precipitation of $CaCO_3$. *Soil Biology and Biochemistry*, 31:1563–1571, 1999. doi: 10.1016/S0038-0717(99)00082-6.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

G. Strang and L. Freund. Introduction to applied mathematics. *Journal of Applied Mechanics*, 53(2):480, 1986.

B. Sudret. Uncertainty propagation and sensitivity analysis in mechanical models–contributions to structural reliability and stochastic spectral methods. *Habilitationa diriger des recherches, Université Blaise Pascal, Clermont-Ferrand, France*, page 18, 2007.

B. Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7):964–979, 2008.

B. Sudret. Polynomial chaos expansions and stochastic finite element methods. *Risk and Reliability in Geotechnical Engineering*, pages 265–300, 2014.

F. Suliman, H. French, L. Haugen, and A. Søvik. Change in flow and transport patterns in horizontal subsurface flow constructed wetlands as a result of biological growth. *Ecological Engineering*, 27(2):124–133, 2006.

R. L. Taylor. Feap-a finite element analysis program, 2014.

B. H. Thacker and T. L. Paez. A simple probabilistic validation metric for the comparison of uncertain model and test results. In *16th AIAA Non-Deterministic Approaches Conference*, page 0121, 2014.

E. Thonhofer, E. Luchini, A. Kuhn, and S. Jakubek. Online parameter estimation for a flexible, adaptive traffic network simulation. In *2014 international conference on connected vehicles and expo (ICCVE)*, pages 937–938. IEEE, 2014.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244, 2001.

M. E. Tipping, A. C. Faul, et al. Fast marginal likelihood maximization for sparse Bayesian models. In *AISTATS*, 2003.

L. Troldborg, J. C. Refsgaard, K. H. Jensen, and P. Engesgaard. The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. *Hydrogeology Journal*, 15(5):843–860, 2007.

J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

R. Trotta. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49(2):71–104, 2008.

T. G. Trucano, L. P. Swiler, T. Igusa, W. L. Oberkampf, and M. Pilch. Calibration, validation, and sensitivity analysis: What's what. *Reliability Engineering and System Safety*, 91(10-11):1331–1357, 2006. ISSN 09518320. doi: 10.1016/j.ress.2005.11.031.

C. J. Turner, R. H. Crawford, and M. I. Campbell. Multidimensional sequential sampling for nurbs-based metamodel development. *Engineering with Computers*, 23(3): 155–174, 2007.

A. Urbina, T. L. Paez, A. Urbina, T. Hasselman, W. Wathugala, and K. Yap. Assessment of model accuracy relative to stochastic system behavior. In *44 th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, 2003.

W. C. Van Beers and J. P. Kleijnen. Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *European journal of operational research*, 186(3):1099–1113, 2008.

L. Van Der Maaten, E. Postma, J. Van den Herik, et al. Dimensionality reduction: a comparative. *Journal of Machine Learning Research*, 10(66-71):13, 2009.

H. A. Van der Vorst. Bi-cgstab: A fast and smoothly converging variant of bi-cg for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 13(2):631–644, 1992.

L. A. van Paassen, R. Ghose, T. J. M. van der Linden, W. R. L. van der Star, and M. C. M. van Loosdrecht. Quantifying Biomediated Ground Improvement by

Ureolysis: Large-Scale Biogrout Experiment. *Journal of Geotechnical and Geoenvironmental Engineering*, 136(12):1721–1728, DEC 2010. ISSN 1090-0241. doi: 10.1061/(ASCE)GT.1943-5606.0000382.

W. K. van Wijngaarden, L. A. van Paassen, F. J. Vermolen, G. A. M. van Meurs, and C. Vuik. A reactive transport model for Biogrout compared to experimental data. *Transport in Porous Media*, 111(3):627–648, Feb 2016. ISSN 1573-1634. doi: 10.1007/s11242-015-0615-5.

J. Venna et al. *Dimensionality reduction for visual exploration of similarity structures*. Helsinki University of Technology, 2007.

A. Verruijt. *Theory of Consolidation*, pages 65–90. Springer Netherlands, Dordrecht, 2010. ISBN 978-90-481-3441-0. doi: 10.1007/978-90-481-3441-0_4. URL `https://doi.org/10.1007/978-90-481-3441-0_4`.

F. A. Viana, R. T. Haftka, and V. Steffen. Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. *Structural and Multidisciplinary Optimization*, 39(4):439–457, 2009.

M. Vohra, P. Nath, S. Mahadevan, and Y.-T. T. Lee. Fast surrogate modeling using dimensionality reduction in model inputs and field output: Application to additive manufacturing. *Reliability Engineering & System Safety*, 201:106986, 2020.

E.-J. Wagenmakers, M. Marsman, T. Jamil, A. Ly, J. Verhagen, J. Love, R. Selker, Q. F. Gronau, M. Šmíra, S. Epskamp, et al. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25(1):35–57, 2018.

P.-R. Wagner, R. Fahrni, M. Klippel, A. Frangi, and B. Sudret. Bayesian calibration and sensitivity analysis of heat transfer models for fire insulation panels. *Engineering Structures*, 205(#110063), 2020. doi: 10.1016/j.engstruct.2019.110063. URL `https://doi.org/10.1016/j.engstruct.2019.110063`.

P.-R. Wagner, J. Nagel, S. Marelli, and B. Sudret. UQLab user manual – Bayesian inversion for model calibration and validation. Technical report, Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland, 2021. Report # UQLab-V1.4-113.

L. Wasserman. Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

K. Weishaupt, V. Joekar-Niasar, and R. Helmig. An efficient coupling of free flow and porous media flow using the pore-network modeling approach. *Journal of Computational Physics: X*, 1:100011, 2019.

K. Weishaupt, A. Terzis, I. Zarikos, G. Yang, B. Flemisch, D. A. M. de Winter, and R. Helmig. A hybrid-dimensional coupled pore-network/free-flow model including pore-scale slip and its application to a micromodel experiment. *Transport in Porous Media*, 135(1):243–270, 2020. ISSN 0169-3913.

V. S. Whiffin, L. a. van Paassen, and M. P. Harkes. Microbial carbonate precipitation as a soil improvement technique. *Geomicrobiology Journal*, 24(5):417–423, Aug. 2007. ISSN 0149-0451.

D. C. Wicaksono. Bayesian uncertainty quantification of physical models in thermal-hydraulics system codes. Technical report, EPFL, 2018.

R. R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

M. A. Wilson, E. S. Iversen, M. A. Clyde, S. C. Schmidler, and J. M. Schildkraut. Bayesian model search and multilevel inference for snp association studies. *The Annals of Applied Statistics*, 4(3):1342, 2010.

Y. Xiong, W. Chen, K.-L. Tsui, and D. W. Apley. A better understanding of model updating strategies in validating engineering models. *Computer Methods in Applied Mechanics and Engineering*, 198(15-16):1327–1337, 2009.

D. Xiu. Fast numerical methods for stochastic computations: a review. *Communications in Computational Physics*, 5(2-4):242–272, 2009.

D. Xiu and G. E. Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002.

D. Xu, G. Bisht, K. Sargsyan, C. Liao, and L. R. Leung. Using a surrogate-assisted Bayesian framework to calibrate the runoff-generation scheme in the energy exascale earth system model (e3sm) v1. *Geoscientific Model Development (Online)*, 15(12), 7 2022. doi: 10.5194/gmd-15-5021-2022.

T. Xu and A. J. Valocchi. A Bayesian approach to improved calibration and prediction of groundwater models with structural error. *Water Resources Research*, 51(11): 9290–9311, 2015.

T. Xu, E. Sonnenthal, N. Spycher, and K. Pruess. 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. *Computers and Geosciences*, 32(2):145–165, 2006. ISSN 00983004. doi: 10.1016/j.cageo.2005.06.014.

Y. Yang, J. Chu, B. Cao, H. Liu, and L. Cheng. Biocementation of soil using non-sterile enriched urease-producing bacteria from activated sludge. *Journal of Cleaner Production*, 262:121315, 2020. ISSN 0959-6526. doi: 10.1016/j.jclepro.2020.121315.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007, 2018.

H. N. Yoon, L. Marshall, A. Sharma, and S. Kim. Bayesian model calibration using surrogate streamflow in ungauged catchments. *Water Resources Research*, 58(1): e2021WR031287, 2022.

G. A. Zampogna and A. Bottaro. Fluid flow over and through a regular bundle of rigid fibres. *Journal of Fluid Mechanics*, 792:5–35, 2016.

R. Zhang and S. Mahadevan. Bayesian methodology for reliability model acceptance. *Reliability Engineering & System Safety*, 80(1):95–103, 2003.

L. Zhao, Z. Lu, W. Yun, and W. Wang. Validation metric based on Mahalanobis distance for models with multiple correlated responses. *Reliability Engineering & System Safety*, 159:80–89, 2017.

Y. Zhou, Z. Lu, K. Cheng, and C. Ling. An efficient and robust adaptive sampling method for polynomial chaos expansion in sparse Bayesian learning framework. *Computer Methods in Applied Mechanics and Engineering*, 352:654–674, 2019.

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

Direktoren
Prof. Dr. rer. nat. Dr.-Ing. András Bárdossy
Prof. Dr.-Ing. Rainer Helmig
Prof. Dr.-Ing. Wolfgang Nowak
Prof. Dr.-Ing. Silke Wieprecht

Emeriti
Prof. Dr.-Ing. habil. Dr.-Ing. E.h. Jürgen Giesecke
Prof. Dr.h.c. Dr.-Ing. E.h. Helmut Kobus, PhD

Lehrstuhl für Wasserbau und Wassermengenwirtschaft
Leiterin: Prof. Dr.-Ing. Silke Wieprecht
Stellv.: Dr.-Ing. Kristina Terheiden
Versuchsanstalt für Wasserbau
Leiter: Stefan Haun, PhD

Lehrstuhl für Hydromechanik und Hydrosystemmodellierung
Leiter: Prof. Dr.-Ing. Rainer Helmig
Stellv.: apl. Prof. Dr.-Ing. Holger Class

Lehrstuhl für Hydrologie und Geohydrologie
Leiter: Prof. Dr. rer. nat. Dr.-Ing. András Bárdossy
Stellv.: Dr. rer. nat. Jochen Seidel
Hydrogeophysik der Vadosen Zone
(mit Forschungszentrum Jülich)
Leiter: Prof. Dr. J.A. Sander Huisman

Lehrstuhl für Stochastische Simulation und Sicherheitsforschung für Hydrosysteme
Leiter: Prof. Dr.-Ing. Wolfgang Nowak
Stellv.: apl. Prof. Dr.-Ing. Sergey Oladyshkin

***VEGAS,* Versuchseinrichtung zur Grundwasser- und Altlastensanierung**
Leiter: Dr.-Ing. Simon Kleinknecht
PD Dr.-Ing. Claus Haslauer

Verzeichnis der Mitteilungshefte

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 Parameteridenti-fikation*, 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

162 Kebede Gurmessa, Tesfaye: *Numerical Investigation on Flow and Transport Characteristics to Improve Long-Term Simulation of Reservoir Sedimentation*, 2007, ISBN 3-933761-66-2

163 Trifković, Aleksandar: *Multi-objective and Risk-based Modelling Methodology for Planning, Design and Operation of Water Supply Systems*, 2007, ISBN 3-933761-67-0

164 Götzinger, Jens: *Distributed Conceptual Hydrological Modelling - Simulation of Climate, Land Use Change Impact and Uncertainty Analysis*, 2007, ISBN 3-933761-68-9

165 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Stuhrmann, Matthias: *VEGAS – Kolloquium 2007,* Tagungsband zur Veranstaltung am 26. September 2007 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2007, ISBN 3-933761-69-7

166 Freeman, Beau: *Modernization Criteria Assessment for Water Resources Planning; Klamath Irrigation Project, U.S.*, 2008, ISBN 3-933761-70-0

167 Dreher, Thomas: *Selektive Sedimentation von Feinstschwebstoffen in Wechselwirkung mit wandnahen turbulenten Strömungsbedingungen*, 2008, ISBN 3-933761-71-9

168 Yang, Wei: *Discrete-Continuous Downscaling Model for Generating Daily Precipitation Time Series*, 2008, ISBN 3-933761-72-7

169 Kopecki, Ianina: *Calculational Approach to FST-Hemispheres for Multiparametrical Benthos Habitat Modelling,* 2008*,* ISBN 3-933761-73-5

170 Brommundt, Jürgen: *Stochastische Generierung räumlich zusammenhängender Niederschlagszeitreihen*, 2008, ISBN 3-933761-74-3

171 Papafotiou, *Alexandros: Numerical Investigations of the Role of Hysteresis in Heterogeneous Two-Phase Flow Systems*, 2008, ISBN 3-933761-75-1

172 He, Yi: *Application of a Non-Parametric Classification Scheme to Catchment Hydrology*, 2008, ISBN 978-3-933761-76-7

173 Wagner, Sven: *Water Balance in a Poorly Gauged Basin in West Africa Using Atmospheric Modelling and Remote Sensing Information*, 2008, ISBN 978-3-933761-77-4

174 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Stuhrmann, Matthias; Schrenk, Volker: *VEGAS-Kolloquium 2008 Ressource Fläche III*, Tagungsband zur Veranstaltung am 01. Oktober 2008 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2008, ISBN 978-3-933761-78-1

175 Patil, Sachin: *Regionalization of an Event Based Nash Cascade Model for Flood Predictions in Ungauged Basins*, 2008, ISBN 978-3-933761-79-8

176 Assteerawatt, Anongnart: *Flow and Transport Modelling of Fractured Aquifers based on a Geostatistical Approach,* 2008, ISBN 978-3-933761-80-4

177 Karnahl, Joachim Alexander: *2D numerische Modellierung von multifraktionalem Schwebstoff- und Schadstofftransport in Flüssen*, 2008, ISBN 978-3-933761-81-1

178 Hiester, Uwe: *Technologieentwicklung zur In-situ-Sanierung der ungesättigten Bodenzone mit festen Wärmequellen,* 2009, ISBN 978-3-933761-82-8

179 Laux, Patrick: *Statistical Modeling of Precipitation for Agricultural Planning in the Volta Basin of West Africa,* 2009, ISBN 978-3-933761-83-5

180 Ehsan, Saqib: *Evaluation of Life Safety Risks Related to Severe Flooding,* 2009, ISBN 978-3-933761-84-2

181 Prohaska, Sandra: *Development and Application of a 1D Multi-Strip Fine Sediment Transport Model for Regulated Rivers,* 2009, ISBN 978-3-933761-85-9

182 Kopp, Andreas: *Evaluation of CO_2 Injection Processes in Geological Formations for Site Screening,* 2009, ISBN 978-3-933761-86-6

183 Ebigbo, Anozie: *Modelling of biofilm growth and its influence on CO_2 and water (two-*

phase) flow in porous media, 2009, ISBN 978-3-933761-87-3

184 Freiboth, Sandra: *A phenomenological model for the numerical simulation of multiphase multicomponent processes considering structural alterations of porous media*, 2009, ISBN 978-3-933761-88-0

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 ur-ban 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 Geo-statistical 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

291 Moreno Leiva, Simón: *Optimal planning of water and renewable energy systems for copper production processes with sector coupling and demand flexibility*, 2022, ISBN 978-3-942036-95-5

292 Schönau, Steffen: *Modellierung von Bodenerosion und Sedimentaustrag bei Hochwasserereignissen am Beispiel des Einzugsgebiets der Rems*, 2022, ISBN 978-3-942036-96-2

293 Glatz, Kumiko: *Upscaling of Nanoparticle Transport in Porous Media,* 2022, ISBN 978-3-942036-97-9

294 Pavía Santolamazza, Daniela: *Event-based flood estimation using a random forest algorithm for the regionalization in small catchments*, 2022, ISBN 978-3-942036-98-6

295 Haun, Stefan: *Advanced Methods for a Sustainable Sediment Management of Reservoirs*, 2022, ISBN 978-3-942036-99-3

296 Herma, Felix: *Data Processing and Model Choice for Flood Prediction*, 2022, ISBN 978-3-910293-00-7

297 Weinhardt, Felix: *Porosity and permeability alterations in processes of biomineralization*

in porous media - microfluidic investigations and their interpretation, 2022, ISBN 978-3-910293-01-4

298 Sadid, Najibullah: *Bedload Transport Estimation in Mountainous Intermittent Rivers and Streams*, 2023, ISBN 978-3-910293-02-1

299 Mohammadi, Farid: *A Surrogate-Assisted Bayesian Framework for Uncertainty-Aware Validation Benchmarks*, 2023, ISBN 978-3-910293-03-8

Die Mitteilungshefte ab der Nr. 134 (Jg. 2005) stehen als pdf-Datei über die Homepage des Instituts: www.iws.uni-stuttgart.de zur Verfügung.