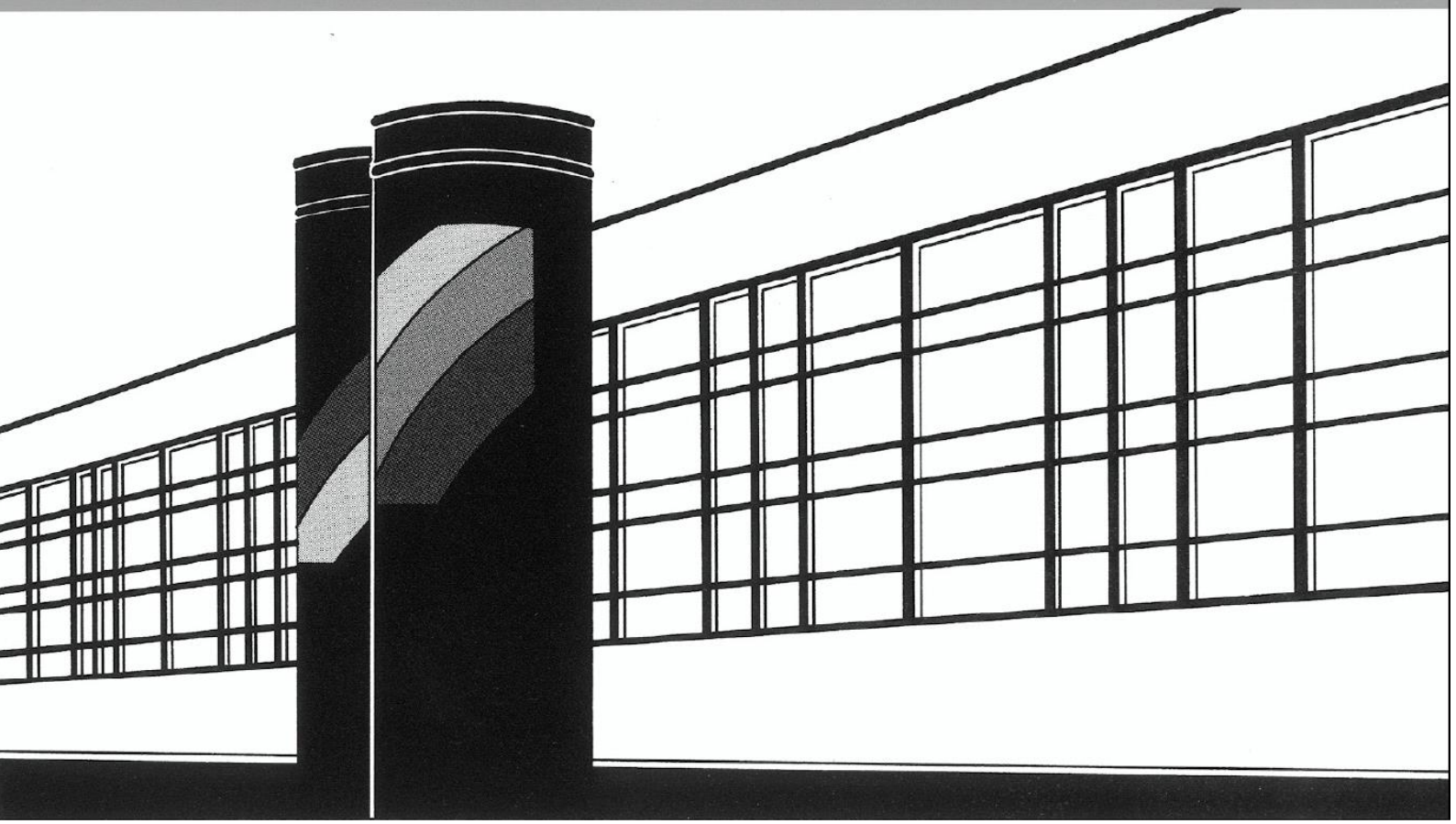
Universität Stuttgart

Institut für Wasser- und Umweltsystemmodellierung

Mitteilungen



Heft 238 Andreas Geiges

Efficient concepts for optimal experimental design in nonlinear environmental systems

Efficient concepts for optimal experimental design in nonlinear environmental systems

Von der Fakultät Bau- und Umweltingenieurwissenschaften der Universität Stuttgart
und dem Stuttgart Research Centre for Simulation Technology
zur Erlangung der Würde eines Doktors der
Ingenieurwissenschaften (Dr.-Ing.) genehmigte Abhandlung

Vorgelegt von

Andreas Geiges

aus Gengenbach

Hauptberichter:	Jun-Prof. Dr.-Ing. Wolfgang Nowak
Mitberichter:	Prof. Dr. rer.nat. Dr.-Ing. Andràs Bárdossy
	Prof. Yoram Rubin
Tag der mündlichen Prüfung:	25.11.2014

Institut für Wasser- und Umweltsystemmodellierung der Universität Stuttgart

2014

Heft 238

Efficient concepts for optimal experimental design in nonlinear environmental systems

von
Dr.-Ing.
Andreas Geiges

Eigenverlag des Instituts für Wasser- und Umweltsystemmodellierung
der Universität Stuttgart

D93 Efficient concepts for optimal experimental design in nonlinear environmental systems

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

Andreas Geiges:
Efficient concepts for optimal experimental design in nonlinear environmental systems, Universität Stuttgart. - Stuttgart: Institut für Wasser- und Umweltsystemmodellierung, 2014

(Mitteilungen Institut für Wasser- und Umweltsystemmodellierung, Universität Stuttgart: H. 238)
Zugl.: Stuttgart, Univ., Diss., 2014
ISBN 978-3-942036-42-9

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

Herausgegeben 2014 vom Eigenverlag des Instituts für Wasser- und Umweltsystem-modellierung

Druck: Document Center S. Kästl, Ostfildern

Acknowledgements

At first, I want to thank my supervisor Wolfgang Nowak, who always had an open door, indicating that there was always time for discussions. It was very special that despite all his time constrains, discussions usually exceeded the requested five minutes by far, since Wolfgang kept discussing. His effort and individual support made this thesis possible. Along the way he somehow managed to grow my interest in the field of statistics. At this point I even like to pursue the topic in my future career, for which I am not sure to thank or curse him.

Almost as import were my colleagues, who were essential to ask those stupid questions, which are said not to exist. I had the pleasure see the starting of a small group of potential scientist, exiled in a wooden barrack, to see the group grow and finally move in an excellent fortress of science, made of concrete. I want to thank specially Phillip, Jonas and Julian for the gazillion discussions, which were needed to sort my wandering thoughts, and for the constant attempts to fight the routine problems that occur when living in a futuristic concrete working environment. Together with a lot others, Lena helped me at various lunch breaks to overcome seldom droughts in the dessert of stochastics.

Important support and ideas came from my co-advisors Andràs Bárdossy and Yoram Rubin, for which I am very grateful.

The possibility to work on parts for my thesis at the University of California in Berkeley was an overwhelming experience. For this, I want to thank Yoram Rubin, who made this possible, welcomed me in his work group and guided my progress in various discussions. The time in Berkeley was unique and allowed me to make many important experiences and valuable new friends.

Finally, I want to thank my family who supported me any time and my friends who made my time in Stuttgart fun and enjoyable. Special thanks to those who joined me in climbing, where I found another kind of challenge and a useful distraction from the theoretic workday problems.

At last, I want to thank the Deutsche Forschungsgemeinschaft (DFG) and the Stuttgart Research Center for Simulation Technology (SimTech) for financing the related research project.

Contents

List of Figures

List of Tables

Notation

Symbol	Definition	Units
Greek letters		
α	dispersivity	[L]
Γ	boundary of the model domain	[-]
$\Gamma()$	Gamma function	
$\boldsymbol{\varepsilon}_{(\cdot)}$	measurement error of the quantity denoted by subscript	
$\boldsymbol{\varepsilon}^m$	model error	
$\boldsymbol{\theta}$	set of structural parameters	
κ	Matérn shape parameter	[-]
λ	correlation length	[L]
μ	arithmetic mean	
ν	ambient flow angle	[rad]
$\boldsymbol{\xi}$	conceptual model parameter	
ψ	rejection factor	[-]
Φ	design utility function	
ϕ	data utility function	
Latin letters		
c	contaminant concentration	[M L^{-3}]
d	drawdown	[L]
$\mathbf{d}$	set of decision variables	
$\mathbf{d}_{opt}$	set of optimal decision variables	
$\mathbf{D}$	design space	
$\mathbf{D}_p$	dispersion tensor	[L^2 T^{-1}]
$f()$	function	
$\mathbf{f}_{[\cdot]}$	model function for quantity denoted by subscript	
F	cumulative density function	
$\mathcal{F}$	Fisher information matrix	
$\mathbf{H}$	linear sensitivity matrix	
H	discrete entropy	[nats]
h	hydraulic head	[L]
h	differential entropy	[nats]
h_{rel}	relative entropy	[nats]
$\mathbf{k}$	set of discrete model choice parameters	
K	kernel function	

$\mathbf{K}_h$	hydraulic conductivity	[L T^{-}1]
L	likelihood function	
$\dot{m}$	contaminant mass flux	[M T^{-1} L^{-2}]
M	stochastic model representation	
$n.$	size of variable denoted by subscript	
p	probability function	
$\mathbf{q}$	flux / Darcy velocity	[LT^{-1}]
$\mathbf{Q}.$	covariance matrix	
r	separation distance	[L]
$\boldsymbol{R_\varepsilon}$	error covariance matrix	
$\boldsymbol{R_\varepsilon^m}$	model error covariance matrix	
$\mathbf{s}$	model parameters	
$\mathbf{S}$	augmented set of uncertain model parameters	
S	saturation	[-]
T	transmissivity	[L^2 T^{-1}]
t	time	[T]
$\mathbf{v}$	seepage velocity	[L T^{-1}]
$\mathbf{w}$	weight vector posterior to measurement	
$\mathbf{W}$	weight matrix prior to measurement	
$\mathbf{y}$	set of potential measurement values	
$\mathbf{y}_0$	hypothetical measurement values	
$\mathbf{z}$	set of model predictions (output)	

Operators and Distributions

δ	Kronecker operator
$\partial()$	partial derivative
$C()$	covariance function
$E_{[\cdot]}()$	expectation operator
∇	nabla operator
$\mathcal{N}$	normal distribution
$\mathcal{U}$	uniform distribution
$V_{[\cdot]}()$	variance operator

Abbreviations

AESS	Average effective sample size
BF	Bootstrap filter
BMA	Bayesian model averaging
ESS	Effective sample size
ETA	evapotranspiration
GA	Genetic algorithm
GS	Greedy search
KL	Kullback-Leibler divergence
LAI	Leaf area index
MI	Mutual information
NRMSE	Normalized root mean squared error
STD	Standard deviation
pdf	probability density function

Abstract

In modern scientific and practical applications, complex simulation tools are increasingly used to support various decision processes. Growing computer power in the recent decades allowed these simulation tools to grow in their complexity in order to model the relevant physical processes more accurately. The number of required model parameter that need to be calibrated is strongly connected to this complexity and hence is growing as well.

In environmental systems, in contrast to technical systems, the relevant data and information for adequate calibration of these model parameters are usually sparse or unavailable. This hinders an exact determination of model parameters, initial and boundary conditions or even the correct formulation of a model concept. In such cases, stochastic simulation approaches allow to proceed with uncertain or unknown parameters and to transfer this input uncertainty to estimate the prediction uncertainty of the model. Thus, the predictive quality of an uncertain model can be assessed and thus represents the current state of knowledge about the model prediction.

In the case that the prediction quality is judged to be insufficient, new measurement data or information about the real system is required to improve the model. The high costs for taking measurements in the subsurface and their differing information value make optimized data acquisition campaigns indispensable. For maximizing the benefits of such campaigns, it is necessary to assess the expected data impact of measurements that are collected according to a proposed campaign design. This allows to identify the so called 'optimal design' that promises the highest expected data impact with respect to the particular model purpose.

This thesis addresses data impact analysis of measurements within nonlinear systems or nonlinear parameter estimation problems. In contrast to linear systems, data impact in nonlinear systems depends on the actual future measurement values, which are unknown at the stage of campaign planing. For this reason, only an expected value of data impact can be estimated, by averaging over many potential sets of future measurement values. This nonlinear analysis repeatedly employs nonlinear inference methods and is therefore much more computationally cumbersome than linear estimates.

Therefore, the overall purpose of this thesis is to develop new and more efficient methods for nonlinear data impact analysis, which allow tackling complex and realistic applications for which in the past only linear(ized) methods were applicable. This thesis separated efficiency of data impact estimation into three different facets:

- **Accuracy**: Accurate estimates of data impact for nonlinear problems demand for rigorous and linearization-free techniques. They also require lean and efficient implementations allowing for adequate statistical discretization at any complexity level. Pursuing the goal

of accuracy requires answering the following research question:
Which available tools in the literature allow a flexible and accurate data impact estimation in nonlinear systems, how can they be combined most effectively and how can they be improved?

- **Computational speed**: The second part of this thesis identifies speed-up potentials within nonlinear data impact estimation, which neither introduce simplifications nor lead to less accuracy. Especially in the context of optimization of data collection, such an analysis is evaluated for thousands of proposed design candidates. Thus, the individual required computation time for a single data impact analysis is a key factor for reducing the overall required computing time of the entire optimization process. Therefore, data impact estimation is reviewed within an information-theoretic background to identify possible speedup strategies and to address the following question:
Which *theoretical* potential can be identified to accelerate nonlinear data impact estimation without using further approximations and how well can it be exploited in *practice*?

- **Robustness**: Model-based data impact estimation is affected by model uncertainty just as any other model output. In fact, the prediction of the actual data impact is impossible, as this would require knowledge about the future measurement values and hence perfect models with perfect model parameters. Therefore, the uncertain model is as well employed to simulate future measurement values. This additional usage introduces even stronger dependencies on the prior model. One way to improve the robustness of the data impact is connecting acquisition design and its execution interactively. In this case, I hypothesize that the model improves step by step due to new available data, which leads to more accurate and robust data impact analysis and therefore to superior data acquisition designs. Considering such interactive schemes requires to answer the following question:
Can interactive mechanisms be introduced to increase the robustness of nonlinear data impact estimation related to the uncertain system model and how much is the data impact improved by this interaction?

The approach taken in this thesis for achieving the defined goals can be divided into three consecutive steps, which coincide with the tackling of the three questions above:

First part - Accuracy: The first goal of this thesis is the development of a nonlinear and fully flexible reference framework for the accurate estimation of data impact. The core of the developed method is the bootstrap filter, which was identified as the most efficient method for fast and accurate simulation of repeated of nonlinear Bayesian inference for many potential future measurement values. The method is implemented in a strict and rigorous Monte-Carlo framework based on a pre-computed ensemble of model evaluations. The non-intrusive nature of the framework allows its application for arbitrary physical systems and the consideration of any type of uncertainty.

The need for nonlinear data impact analysis and their advantages versus linear methods are shown in several test cases. In all considered test cases, data acquisition designs that are optimized with respect to their nonlinear data impact show substantially better results than their

linearly designed counterparts. One particular example addresses the importance of the accurate representation of the current model uncertainty on every conceptual level, and how much this uncertainty affect the final data collection design. A second example from crop plant growth modeling shows how different conceptual models can be considered in the design process as well. The existence of competing model alternatives is implemented with the concepts of Bayesian model averaging and included in the framework at no additional costs.

Second part - Computational speed: The flexibility to consider any type of uncertainty and the assessment of nonlinear data impact requires substantially more computational time compared to linearized methods. For this reason, the second part of this thesis investigates the theoretical background of data impact analysis in order to identify potentials to speed up this analysis. The key idea followed in this part originates from the well-known symmetry of Bayes Theorem and of a related information measure called Mutual Information. Both allow considering a reversal of the direction of information analysis, in which the roles of potential measurement data and the relevant model prediction are exchanged. Since the space of potential measurements is usually much larger than the space of model prediction values and since both have fundamentally different properties, the reversal of the information assessment offers a high potential for increasing the evaluation speed.

The actual implementation of such a reverse formulation is tested in a numerical study. The goal is to show to which degree the theoretic speed up can be transferred to practical applications. The study shows that the actual speedup is potentially high, but heavily depends on the particular estimation technique, which is used for probability density distributions. This is especially the case for high-dimensional designs that comprise of larger numbers of measurements. The tested estimators allowed evaluation times that are up to two orders of magnitudes faster than the fastest classical analysis.

In addition, the general idea of reversing information assessment allows the development of an approximation of data impact that relies on parametric entropy estimation. This partly linearized estimate only requires evaluation times that are comparable to fully linear methods, but offers much higher approximation quality with respect to the fully nonlinear reference estimate.

Third part- Robustness: The last basic facet of an efficient data impact estimation considers the robustness of such estimates with regard to the uncertainty of the underlying model. Basically, model-based data impact estimates are subject to the same uncertainty as any other model output. Thus, the data impact estimate can be regarded as just another uncertain model prediction. Therefore, the high uncertainty of the model (which is the reason for the search for new calibration data) also affects the process of evaluating the most useful new data.

This dependency on the initial uncertainty can not be overcome without incorporating new data. Therefore, the pursued approach in this part aims to increase the overall robustness using an interactive design approach. It adapts the ongoing data acquisition campaign based on newly available measurement data. Thus, later data collection in such a sequential interactive scheme is based on an improved state of knowledge and is more specific to the real system.

This requires the design framework to be sufficiently fast to be executed parallel to the ongoing data acquisition.

A synthetic case study proves the success for two actual implementations of interactive design strategies. The study reveals the major effect of the data within the design process and simultaneously indicates that this effect is much more important than a globally optimized data acquisition design. A sequential non-global optimization strategy, which interactively incorporated the data as early as possible, was found to be most robust and therefore performed best in the given scenario.

Synergies: Finally, a small concluding study demonstrates the synergy effects between the methods from the three previous parts of this thesis. The combination of interactive design strategies (which favor sequentially executed designs of low dimensionality) and reverse-like data impact formulations (which show the highest speedup effect for low dimensions as well) allows considering optimization of data acquisition campaign in real-time.

Conclusion: In summary, the developed methods and theoretic principles allow for more efficient evaluation of nonlinear data impact. The use of nonlinear measures for data impact lead to an essential improvement of the resulting data acquisition design with respect to a relevant prediction quality. These methods are flexibly applicable for any physical model system and allow the consideration of any degree of statistical dependency. Especially the interactive approach that counters the high initial uncertainties of the model does lead to huge improvement in the design of data acquisition. All achieved conceptual and practical improvements in the evaluation of nonlinear data impact assessment allow using such powerful nonlinear methods also for complex and realistic problems.

Kurzfassung

In der modernen Wissenschaft, sowie in praktischen Anwendungen, bilden Simulationsmodelle in immer größerem Maße die Grundlage für komplexe Entscheidungsprozesse. Die Komplexität solcher Simulationsmodelle steigt stetig mit der verfügbaren Rechenleistung, mit dem Ziel die komplexen Prozesse im realen System bestmöglich wiederzugeben. Damit erhöht sich gleichzeitig die Anzahl der Modellparameter, welche mit Hilfe von Messdaten oder anderen Informationen über das reale System kalibriert werden müssen.

Für Umweltsysteme sind, im Gegensatz zu technischen Anwendungen, die nötigen Messdaten für die adäquate Bestimmung von Modellparametern, Rand- und Initialbedingungen und für Entscheidungen über das korrekte konzeptionelle Modell häufig nicht ausreichend verfügbar. Daher werden stochastische Simulationsmethoden angewendet um unsichere oder unbekannte Parameter in numerischen Simulation handhaben zu können. Solche stochastischen Verfahren erlauben es ebenfalls, von der Unsicherheit der Eingangsparameters auf die Vorhersageunsicherheit des Modells zu schließen.

Im Falle einer zu großen Vorhersageunsicherheit werden zusätzliche Messdaten aus dem modellierten System benötigt. Messungen in Umweltsystemen und speziell im Untergrund sind sehr teuer und erfordern oft aufwändige Messtechnik. Direkte Methoden liefern lediglich lokale Informationen, welche zusätzlich durch Störungen bei der Beprobung fehlerbehaftet sein können. Im Gegensatz dazu erlauben indirekte Messmethoden lediglich Abschätzungen räumlich gemittelter Größen. Daher ist es besonders in diesem Zusammenhang notwendig, den (Informations-)Wert solcher (fehlerbehafteter) Messdaten im Voraus abzuschätzen. Dieser Informationswert potentieller zukünftiger Messdaten ist jeweils von der relevanten Modelvorhersage und damit von der Zielstellung abhängig und muss problemspezifisch definiert werden.

Für die Planung einer effizienten und kostenoptimierten Messkampagne ist es ebenfalls notwendig, den Datennutzwert zukünftiger Messdaten abzuschätzen. Diese Doktorarbeit beschäftigt sich speziell mit der Datenwertanalyse für nichtlineare Probleme. Im Gegensatz zu linearen Problemen, ist der Datenwert für nichtlinear Zusammenhänge von dem zukünftigen (und daher unbekannten) Messwert anhänig. Der Nutzwert kann daher nur als Erwartungswert über alle möglichen Messwerte abgeschätzt werden. Diese Abschätzung erfordert nichtlineare Inferenzmethoden für jeden potenziellen Messwert und ist deshalb, im Gegensatz zu linearen Schätzern, rechnerisch sehr aufwändig.

Das Ziel dieser Arbeit ist es neue effiziente Methoden für die nichtlineare Datenwertanalyse zu entwickeln, welche eine optimale Datenerhebung in komplexen nichtlinearen Systemen ermöglicht. Diese Arbeit geht ins Besondere auf drei wesentliche Gesichtspunkte ein, welche im weiteren Sinne eine effiziente Datenwertanalyse erlauben:

- **Genauigkeit:** Die Genauigkeit des vorhergesagten Informationswerts einer Messung hängt hauptsächlich von dem Grad der berücksichtigten statistischen Abhängigkeiten ab. Deshalb wird in dieser Arbeit als erster Schritt eine akkurate nichtlineare Schätzmethode entwickelt, welche Abhännigkeiten beliebiger Ordung berücksichtigt. Um einen möglichst großen Anwendungsbereich zu gewährleisten, sollte die entwickelte Methode flexibel und möglichst generell ausgelegt sein. Dies führt zu folgender Fragestellung:
  Welche Methoden zur nichtlinearen statistischen Analyse stehen in der Literatur zur Verfügung und wie können diese erfolgreich kombiniert oder weiterentwickelt werden um eine effiziente nichtlineare Datenwertanalyse zu ermöglichen?

- **Rechenaufwand:** Speziell im Rahmen der Versuchsplanung ist es notwendig, Datenwertanalysen für eine große Zahl verschiedener konkurrierender Messkampagnen auszuwerten. Daher ist die notwendige Rechenzeit bzw. der -Aufwand einer einzelnen Analyse oft das ausschlagende Kriterium, das über die Anwendbarkeit einer Methode entscheidet. Der zweite Teil dieser Arbeit beleuchtet die Theorie hinter der Datenanalyse um folgende Frage zu beantworten:
  Welches theoretische Potential eröffnet sich für die Beschleunigung der Datenwertanalyse und in wieweit lässt sich dieses Potential in der Praxis ausnutzen?

- **Robustheit:** Das Ergebnis einer simulationsgestützten Versuchsplanung hängt maßgeblich von der Qualität des zur Simulation verwendeten Modells ab. Vor der Erhebung neuer Messdaten zur Verbesserung des Models leidet auch der geschätzte Datenwert an minderer Vorhersagequalität. Daher beschäftigte sich der dritte Teil dieser Doktorarbeit mit Möglichkeiten die Robustheit der Versuchsplanung gegenüber der anfangs hohen Modellunsicherheiten zu erhöhen. Neben statistischen Verfahren aus der Literatur hat eine möglichst frühe interaktive Einbeziehung neuer Messdaten in den Planungsprozess das größte Potential die Robustheit insgesamt zu verbessern. In diesem Zusammenhang beantwortet diese Arbeit folgende Fragestellung:
  Welche interaktiven Mechanismen lassen sich in der Versuchsplanung von Messkampagnen einsetzten und inwieweit lässt dich ein verbessertes Gesamtergebnis garantieren?

Die folgende Arbeit entwickelt effiziente Methoden in Hinsicht auf die drei vorgestellten Gesichtspunkte, welche gleichzeitig die Struktur der Arbeit vorgeben:

Erster Teil - Genauigkeit: Dieser Teil der Arbeit entwickelt eine allgemeine und flexibel anwendbare Methode für die nichtlineare Datenwertschätzung. Den Kern der Methode bildet der Bootstrap Filter, welcher wiederholt für nichtlineare Inferenz möglicher Daten genutzt wird. Die vollständige Simulation der Bayesschen Inferenz und deren Auswirkungen auf das Modell ermöglicht eine umfassende nichtlineare Analyse aller statistischen Zusammenhänge zwischen den vorgeschlagenen Messdaten und der mageblichen Modellvorhersage. Die statistischen Abhängigkeiten werden nach dem Monte-Carlo Prinzip und basierend auf einem Ensemble vorberechneter Modellauswertungen approximiert. Das zu Grunde gelegte nicht-intrusive Prinzip lässt sich mit beliebigen Simulationswerkzeugen kombinieren und unabhängig von den physikalischen Eigenschaften des Systems anwenden.

Die Vorteile dieser nichtlinearen Datenanalyse werden in mehreren Beispielen aufgezeigt und mit linearen Methoden verglichen. Messkampagnen zur Datenerhebung, welche auf Basis von nichtlinearer Datenwertanalyse optimiert wurden, zeigen dabei erheblich bessere Resultate. Es wird illustriert wie wichtig die anfängliche Berücksichtigung aller Unsicherheiten für eine sinnvolle Informationsanalyse ist. Dabei zeigt sich der Einfluss von parametrischer, struktureller und konzeptioneller Unsicherheit auf den Wissensbedarf und das daraus resultierende optimierte Messdesign.

Eine weiterführende Studie zeigt die Kombination mit Bayesscher Modellmittelung anhand eines Beispiels der Modellierung von saisonalem Wachstum einer Weizenkultur. Bei dieser Fragestellung kommen mehrere konzeptionelle Modelle zum Einsatz, welche ohne konzeptionellen Aufwand bei der nichtlinearen Datenanalyse berücksichtigt werden können.

Zweiter Teil - Rechenaufwand: Die Flexibilität, jegliche Arten von Unsicherheiten zu berücksichtigen, zusammen mit einer nichtlinearen Analyse erfordern einen sehr hohen Rechenaufwand. Daher beschäftigt sich der zweite Teil dieser Arbeit mit Möglichkeiten die nichtlineare Analyse zu beschleunigen, ohne dabei auf Vereinfachungen zurückzugreifen. Dabei wird die Symmetrie des zu Grunde liegenden Bayes-Theorem für die Inferenz von Daten und ein daraus resultierendes Maß für Information näher betrachtet. Diese Symmetrie erlaubt es in der Daten- oder Informationsanalyse die Rolle von Messdaten und Modellvorhersage zu vertauschen.

Da der Raum potentieller Messdaten normalerweise erheblich größer ist als der Raum aller möglichen Modellvorhersagen, kann eine Umkehrung der Informationsanalyse zu einer erheblichen Beschleunigung der Analyse führen. In einer numerischen Studie wird untersucht, inwieweit sich dieses theoretische Potential in der Praxis umsetzen lässt. Die tatsächliche erreichbare Beschleunigung hängt dabei maßgeblich von den verwendeten Schätzverfahren für die involvierten Wahrscheinlichkeitsdichten ab. Die getesteten Schätzer erlaubten eine Reduktion der nötigen Rechenzeit der Datenwertanalyse um bis zu zwei Größenordnungen bei gleichbleibender Schätzgüte.

Zusätzlich erlaubt die Idee einer Umkehrung der Analyserichtung die Entwicklung einer teilweise linearisierten Abschätzung des Datenwerts, welche auf der gleichen Grundidee basiert. Die entwickelte Approximation zeigt erheblich bessere Schätzgüte im Vergleich zu vollinearen Schätzern bei vergleichbaren Rechenaufwand.

Dritter Teil - Robustheit: Ein grundsätzliches Problem der modellbasierten Datenwertanalyse besteht darin, dass das numerische Modell mit unzureichender Vorhersagequalität gleichzeitig die beste verfügbare Grundlage für die Datenwertanalyse für potentielle Messdaten darstellt. Eine solche Datenwertanalyse und die daraus abgeleitete optimale Messkampagne leidet deshalb unter der gleichen Vorhersageunsicherheit.

Da diese anfängliche Unsicherheit nur durch die Nutzung neuer Daten reduziert werden kann, muss ein möglicher Lösungsansatz sich mit der frühest möglichen Einbeziehung verfügbarer Daten in den Optimierungsprozess für die weitere Messkampagne befassen. Der dritte Teil der

Arbeit untersucht daher interaktive Planungsstrategien, welche in der Lage sind eine laufende Messkampagne adaptiv auf neu verfügbare Messungen anzupassen. Die Messkampagne wird dabei in Sequenzen aufgeteilt, welche nacheinander geplant und ausgeführt werden. Die resultierenden Messdaten jeder Sequenz werden in der darauffolgenden Sequenz berücksichtigt und führen damit zu einer spezifischeren und damit verbesserten Planung der nächsten Sequenz.

Eine numerische Studie belegt den Erfolg einer solchen interaktiven Planungsstrategie und zeigt gleichzeitig, dass der Einfluss interaktiv genutzter Messdaten weitaus wichtiger ist als eine global optimierte Messkampagne. Eine einfache sequentielle Optimierung führte mehrfach zu besseren Ergebnissen als eine global optimierte Kampagne auf der Basis des Anfangsmodells.

Synergien: Eine abschließende Studie über die Synergieeffekte zwischen den drei anfänglich separierten Ansätzen dieser Arbeit legt dar, wie diese erfolgreich kombiniert werden können. Diese Synergien erlauben es problemspezifisch und gezielt den relativ hohen Rechenaufwand für die modell-gestützte Planung von Messkampagnen soweit zu reduzieren, dass eine Planung und Optimierung von Messkampagnen ohne massive Parallelisierung in Echtzeit ermöglicht wird. Gleichzeitig zeigt die Kombination approximierter Schätzer mit einer interaktiven Planungsstrategie die Tendenz verminderter Schätzfehler, wodurch das Gesamtergebnis verbessert wird.

Fazit: Zusammenfassend erlauben die erarbeiteten Methoden und theoretischen Grundlagen dieser Arbeit eine effizientere und schnellere Datenwertanalyse für komplexe Systeme unter Berücksichtigung von nichtlinearen Zusammenhängen und beliebigen Arten von Unsicherheiten. Die erreichte Beschleunigung der Datenwertanalyse erlaubt es, diese auf komplexere und realistischere Problemstellungen anzuwenden. Zusätzlich führen die entwickelten interaktiven Planungsverfahren zu einer beträchtlichen Steigerung des Informationsgewinns in Messkampagnen.

1. Introduction

1.1. Motivation

Any resource use or technical application in the subsurface is hindered by limited knowledge about the system. Especially the fact that the subsurface is hidden to sight, yet is known to be spatially heterogeneous within its properties, will never allow a complete system description [*Oreskes et al.*, 1994]. Furthermore, data acquisition in the subsurface is usually expensive and technically challenging. Therefore, it is most important to optimally design the acquisition of measurement data to maximize their output. Well-guided data acquisition can help to minimize uncertainty in modeling, simulated predictions, planing and resource use. To improve methods for guiding data acquisition is the focus of the present dissertation.

Human activities and exploitation in and on the subsurface have steadily increased over time, and nowadays the subsurface is used for multiple purposes at the same time. These comprise, among many others, waste deposition (waste disposal sites, CO_2 sequestration [e.g., *Kopp et al.*, 2009; *Walter et al.*, 2012], nuclear waste disposal [e.g., *Olivella et al.*, 1994]), energy storage and extraction [e.g., *Harlow and Pracht*, 1972] and enhanced mining techniques, like oil shale extraction [e.g., *Hu et al.*, 1999; *Ogunsola and Berkowitz*, 1995] and the often criticized fracking [e.g., *Myers*, 2012]. Furthermore, the amount of activities have reached a level at which the different activities are affecting each other in an unforeseeable manner.

At the same time, the subsurface is the most important source of clean drinking water. Human activities in and on the subsurface endanger the quality of this resource, which was formerly safely stored from harmful influences and naturally treated in the subsurface. Keeping the subsurface as a reliable drinking water source and mitigating the negative effects of any exploitation requires proper risk assessment and effective monitoring. This includes identifying, assessing and judging future impacts of all past, present and future human activities.

To this end, numerical models for simulating the relevant processes in the subsurface are required for predicting future system states and system responses. However, simulations of complex and dynamic processes over large time scales remain challenging, even with state-of-the-art models and todays computing power. For environmental systems, accurate predictions are even more hindered by the inherent lack of knowledge about the system. Facing ubiquitous uncertainty about environmental systems requires to extend deterministic models towards stochastic modeling frameworks. Such analyses allow the modeler to quantify model uncertainty in general and to supplement model predictions with error bounds and confidence intervals [e.g., *Christakos*, 1992; *Rubin*, 2003]. They require a proper representation of the available data by uncertain model parameters as well as methods to transfer this input uncertainty to the uncertainty of the relevant model output.

In the case that the analyzed uncertainty of an aspired model prediction is too large, additional measurement data are required to revise the model. Hence, efficient methods for designing expensive field campaigns are required for only collecting the most informative data. Such optimal designs should consider all currently available data and knowledge about system processes, such that only the most important complementary information is collected. In fact, Design of Experiments (DoE) [e.g., *Ghosh*, 1999; *Pukelsheim*, 2006; *Plackett and Burman*, 1946] is a framework to rationally design arbitrary (scientific) experimentation tasks and can be applied for the design of data acquisition campaigns. Such an optimized campaign design ensures the highest possible benefits and thus leads to the best allocation of available resources towards reduced model uncertainty, based on the present available data.

Potential data acquisition designs are rated based on their expected data impact, which is an estimate how much relevant information the data will be able to provide related to a given task. The data impact is estimated based on the current state of knowledge, before knowing the resulting measurements values. The quality and complexity of this estimation strongly affects the actual data impact of the selected (allegedly optimal) design and the overall computation time for the optimization. With limited computational power and the requirement of evaluating design trials in the order of hundreds of thousands and more, the majority of the studies in the past used linearized approaches to estimate data impact. However, with increasing model complexity, which often introduces nonlinear dependencies between parameters and model predictions, linearized estimates become inadequate. Thus, there is a great need to make more accurate and efficient DoE methods available to environmental problems that involve nonlinear relations within the corresponding models. In the following, the major shortcomings in the state of the art will be identified.

1.2. State of the art

The design of groundwater sampling has been studied extensively in the past [e.g., *Bogaert and Russo*, 1999; *Christakos and Killam*, 1993; *Loaiciga*, 1989]. Originally, the DoE originated from regression-like problems [e.g., *Box*, 1982; *Federov and Hackl*, 1997; *Pukelsheim*, 2006], was later extended towards the calibration of computer models [e.g., *Gates and Kisiel*, 1974] and was finally linked to the generic design-of-experiments theory [e.g., *Federov and Hackl*, 1997]. For environmental systems, the benefits of proper uncertainty assessment in modeling and decision making is widely recognized [e.g., *Pappenberger and Beven*, 2006; *Rubin*, 2003].

The major source of uncertainty in sub-surface systems stems from the unknown hydraulic conductivity field. It is the main parameter for flow and transport simulation and can vary over several orders of magnitudes. The sparsely available point information is spatially extended by modeling the conductivity field as a space random function (SRF) [*Journel and Huijbregts*, 1978]. SRFs include knowledge about the structural dependencies to extend point information and to limit the allowed parameter space. Specialized designs to optimally estimate spatial parameters or identify structural parameters of a geostatistical model are referred to as geostatistical designs [*Müller*, 2007].

For a particular modeling prediction, general model improvement is unspecific and therefore

rather inefficient in short-therm considerations. For a given task, it is more beneficial to focus directly on the related model prediction or model purpose. As such, data impact is not related primarily to improvements of unknown model parameters, but directly to the model prediction of concern. Shifting the focus in that way implicitly considers only those model parameters as important which are relevant for the improved model prediction [e.g., *Nowak et al.*, 2009]. In the context of task-driven optimal design [e.g., *Ben-Zvi et al.*, 1988; *James and Gorelick*, 1994; *Nowak et al.*, 2009], a relevant output of the numerical model is defined, which is required to be predicted with high accuracy. Prominent examples include remediation designs, which are based on uncertain numerical models [e.g., *Bear and Sun*, 1998; *Copty and Findikakis*, 2000]. In such campaigns, only these measurement data are deemed valuable, which improve the model related to the remediation task. Such a directed improvement of the model towards a particular prediction task allows for more specific, reliable and therefore cheaper remediation designs, without the need to acquire unrealistic amounts of data. In the literature, examples of task-driven design utility can be found such as Value of Information Analysis [e.g., *Howard*, 1966; *Yokota and Thompson*, 2004], data worth analysis which roots back in water resources to *James and Gorelick* [1994], and Bayesian decision analysis, mentioned in *Feyen and Gorelick* [2005].

Engineering approaches often relate to information-theoretic quantities [e.g., *Herrera and Pinder*, 2005], but rarely apply them directly. It can be shown that almost any measure of data impact can be related to a quantity called Mutual Information (MI) [e.g., *Cover and Thomas*, 2006]. Using MI allows to estimate how much information an observation carries about an unknown model parameter or model prediction, while accurately considering any order of dependencies. Information theory arose mainly within the field of computer science over the past decades and was developed for the quantification of information and compression efficiency. It offers a theoretical foundation that may help improving existing engineering approaches by an accurate description of information. *Nowak et al.* [2009] showed that linearized DoE analysis approaches like First-Order-Second Moment (FOSM) [e.g., *Kunstmann et al.*, 2002; *Cirpka and Nowak*, 2004], adjoint-state sensitivities [e.g., *Sykes et al.*, 1985; *Sun*, 1994; *Cirpka and Kitanidis*, 2001] and the static Ensemble Kalman Filter [e.g., *Herrera and Pinder*, 2005] can be linked to Fisher information [e.g., *Edgeworth*, 1908] for linear and multi-Gaussian systems. Unless when applied to strictly linear cases, linearized analyses lose the link to information theory and designs optimized under linear assumptions are only locally optimal [*Federov and Hackl*, 1997, p.100].

In environmental systems, linear dependencies are rather the exception than the rule, for reasons that are highlighted in Sec. 3.3.2. That section provides an extensive list of mechanisms in subsurface hydrology that lead to nonlinear system behavior. An increasing number of studies therefore focus on the nonlinear estimation of data impact. Nonlinear data impact analysis requires expensive averaging over possible future observation values, because data impact depends on the future (and hence unknown) measured value. Therefore, in the past, estimation of nonlinear data impact imposed computational restrictions and only allowed tackling either simple problems or reduced numerical systems.

Furthermore, one has to keep in mind that the model which suffers from high conceptual and parametric uncertainty is employed and analyzed within the estimation of data impact. There-

fore, the estimate of expected data impact can be seen as another prediction quantity of the model that suffers from the same uncertainty. Optimization under uncertainty is a specific challenge within DoE, trying to minimize the negative effects of the uncertainty in the utility function. In hydrology, *Criminisi et al.* [1997] analyzed the robustness of a design facing model uncertainty. Conditional Monte Carlo simulation within the optimization [e.g., *Maxwell et al.*, 1999; *Copty and Findikakis*, 2000] and chance constraints [e.g., *Tung*, 1986; *Bear and Sun*, 1998; *Freeze and Gorelick*, 1999] are used to incorporate the uncertainty into optimization frameworks. For this problem class, a specialized noisy genetic algorithm [*Gopalakrishnan et al.*, 2003] has been designed to effectively optimize an uncertain utility function in an expected sense. There are even multi-objective approaches dealing with uncertainty in optimization [e.g., *Gunawan and Azarm*, 2005]. These include an additional objective to maximize robustness of a design related to the uncertain parameters. All approaches basically assess design robustness with respect to uncertainty and introduce additional safety margins, by sacrificing some of the achievable expected data impact and therefore yield to less variation in the data impact prediction. Therefore, it is a question of risk perception and risk tolerance to choose the design with the highest expected data impact or one which is robust, but inferior in an expected sense.

1.3. Goals

The primary goal of this thesis is to develop new, effective and informed methods for the optimal design of data acquisition campaigns to overcome the shortcomings described in the previous section. Thus, the developed methods should be generally applicable for any system including all complexity levels and be able to face arbitrary sources of uncertainty and arbitrary levels of nonlinearity. With arbitrary sources of uncertainty, parametric uncertainty, structural model uncertainty (e.g., boundary conditions) and even uncertainty in model choice are addressed. The three major goals, addressing the major shortcomings, are:

1. **Accuracy**: Improved accuracy for DoE methods mainly demands for a rigorous and linearization-free estimation of data impact. It also requires lean and efficient implementations to allow for adequate statistical discretization at any level of complexity. Pursuing the goal of accuracy requires to answer the following question:
   Which available tools in the literature allow a flexible and accurate data impact estimation, how can they be combined most effectively and how can they be improved?

2. **Computation speed**: The second goal of this thesis is to identify speed-up potentials within nonlinear data impact estimation that do not rely on simplifications, do not sacrifice some accuracy and do not require additional assumptions. Therefore, data impact estimation is reviewed within an information-theoretic background to identify possible speedup strategies and to address the following question:
   Which *theoretical* potential can be identified to accelerate nonlinear data impact estimation without using further approximations and how well can it be exploited in *practice*?

3. **Robustness**: Model-based data impact estimation in the face of omnipresent model uncertainty is affected by model uncertainty just as any other model prediction. In fact, the

prediction of the actual data impact is impossible, as this would require knowledge about the future measurement values and hence perfect models with perfect model parameters. Therefore, the uncertain model is as well applied to simulate future measurement values. This additional usage introduces even stronger dependencies on the prior model. One way to improve the robustness of the data impact is connecting acquisition design and its execution interactively. When combining design and execution of data acquisition interactively, I hypothesize that the model improves step by step due to newly available data, which leads to more accurate and robust data impact analysis and therefore superior data acquisition designs. Considering such interactive schemes requires to answers the following question:
What interactive mechanisms can be introduced to increase the robustness of nonlinear data impact estimation related to the uncertain system model and how much is the data impact improved by this interaction?

1.4. Approach

The approach taken in this thesis for achieving the the goals defined above can be divided into three consecutive steps:

Step I - Accuracy: First, this thesis extends nonlinear inference techniques towards data impact estimation in an optimal design framework. The chosen approach extends nonlinear inference methods toward the estimation of data impact. This captures the full information content of potential measurements and can consider arbitrary types of statistical dependence that may be relevant for DoE. The implementation in a strict and brute-force Monte-Carlo framework allows for the consideration of any type of data and arbitrary sources of uncertainty. For nonlinear inference problems, data impact depends on the actual (yet) unknown measured value. Therefore the expected data impact is estimated by averaging over many possible future data values. This implementation will serve as an accurate reference that allows the comparison of estimation quality of other implementations and approximations developed later in this thesis and to show the benefits versus linearized approaches.

Step II - Speed: The averaging over many potential conditional model states within nonlinear estimates leads to high computation times and requires large computational resources. Therefore, the second step focuses on concepts to minimize the computational costs of the nonlinear data impact analysis. The information-theoretic background of data impact assessment is reviewed to show that a certain symmetry in Bayes' theorem allows an equivalent reverse formulation of data impact. The key idea is to reverse the direction of information analysis, by swapping the roles of observable data and model prediction in the analysis of statistical dependency. This offers a drastic potential for the reduction of computational costs and, thus, offer possibilities to develop faster implementation techniques. An extensive numerical study will show the massively faster evaluation of nonlinear data impact without using additional approximations. The study will show that the surveyed residuals between both formulations (forward and reverse) merely originate from different implementation techniques and a differ-

ent statistical convergence behavior. In addition, I provide a alternative approximation of data impact that originate from the reverse mindset, but is partly linearized.

Step III - Robustness: The last step in this thesis aims to maximize the robustness of data impact measures in face of uncertainty. To do so, it is necessary to consider the influence of the uncertain model on the current design process. State-of-the-art optimization frameworks under uncertainty cannot overcome the fact that all uncertainty arises due to the yet unobserved system states, and that only new observation data can fundamentally change the model confidence level. Instead, I propose a conceptually different direction to improve the robustness of estimating data impact. My approach is based on the interactive incorporation of data into the design process. This sets aside the traditional separation of the design and execution of a field campaign. The approach allows for dynamic design decisions, which react on newly available data. I propose to generally split the initial design in consecutive sequences and to dynamically adjust the data collection in each sequence based on currently available data. This will lead to the earliest possible use of new data within the design process, decrease the dependency on the prior model, and potentially increase the overall performance.

Synergies: In a last concluding study, the approaches of all three previous steps are combined in a study to specially highlight the arising synergy effects. These effects were not foreseen in advance, but lead to additional potentials to speed up DoE in combination with nonlinear data impact analysis.

Expected benefits: Overall, the techniques for fully nonlinear, efficient and robust estimation of data impact introduced in this thesis will make model-based planing/design of data acquisition accessible to larger, more complex and thus more realistic problems classes. This will help to give reliable model-based decision support for complex practical applications. The increased computation speed will further allow for optimization of field activities and decision support in real time.

1.5. Structure of work

The remainder of this thesis is tightly structured to follow the three steps of the approach. At first, Chaps. 2-3 will provide the necessary theoretical background upon which this thesis is built. Chap. 2 includes the required physical equations from sub-surface hydrology, descriptive and Bayesian statistics, basic principles of information theory and the resulting handling of uncertainty in stochastic modeling. Chap. 3 separately introduces the optimal design theory and data impact estimation, which is mainly built on the theory from the previous chapter. Chapters 4 to 6 are forming the core of this thesis and deal with the central steps of this thesis: Steps I- Nonlinear data impact estimation, Step II - Reverse data impact assessment and Step III - Interactive campaign design. The following Chap. 7 additionally points out synergy effects that manifest from the conclusions within Steps I-III. This integral evaluation is illustrated within a final numerical application. I will summarize and conclude my work in Chap. 8 and give an outlook on possible future work.

2. Basic methods and governing equations

This chapter establishes the fundamental theory and provides all basic methods that are used throughout this thesis. First, I provide the governing equations for the numerical models that are used for application example scenarios. This is followed by an overview of the statistical background applied in this thesis, including descriptive and Bayesian statistics and information theory. In the end, I introduce the Bayesian mindset and the corresponding approaches for handling parametric and structural uncertainty within numerical models.

2.1. Governing physical equations

The majority of application examples in this work are taken from the field of groundwater hydrology. The current section provides the background for numerical modeling of flow and transport processes within the subsurface. One model type comprises flow and transport models for the saturated zone for simulating groundwater pollution processes. A second type of numerical models are heuristic crop plant models that are coupled with a flow and transport model in the unsaturated zone.

2.1.1. Groundwater flow and transport

For the groundwater applications, I consider single-phase flow through the subsurface, modeled as a porous medium. The flow-relevant properties of the porous medium are described by the hydraulic conductivity tensor K_h [LT^{-1}] (here assumed to be locally isotropic for simplicity) and the porosity θ [-]. In the application scenarios used here, groundwater flow is generally considered to be governed by the relations expressed in Darcy's law and by the conservation of mass. The resulting equation after *Bear* [1972] to describe flow in depth-averaged, isotropic aquifers is

$$S\frac{\partial h}{\partial t} + \nabla\cdot(T\nabla h) - Q = 0, \tag{2.1}$$

where S [-] is the specific storage coefficient, h [L] is the hydraulic pressure, T [L^2 T^{-1}] is vertically averaged transmissivity of the aquifer, Q [L/T] is a volumetric depth-integrated source/sink term and t [T] is time. The transmissivity is the integral of the hydraulic conductivity $\mathbf{K}_h$ [L T^{-1}] over the aquifer thickness b [L]. The system border Γ requires boundary conditions that are either defined as fixed values (Dirichlet) for h or fixed flow (Neumann) conditions:

$$\begin{aligned} h &= h_{di} \quad \text{on} \quad \Gamma_{di} \\ -\mathbf{n}\cdot(T\nabla h) &= q_{ne} \quad \text{on} \quad \Gamma_{ne}, \end{aligned} \tag{2.2}$$

where Γ_{di} and Γ_{ne} refer to Dirichlet and Neumann boundaries at which fixed Dirichlet head h_{di} [L] or the respective fixed Neumann fluxes q_{ne} [LT^{-1}] are imposed. In general, $\mathbf{n}$ is a vector normal to the boundary Γ, pointing outwards. No-flow conditions are a special case of the Neumann boundary condition with $q_{ne} = 0$.

For measurements of steady-state pumping tests, the drawdown s [L] for a constant pumping rate is defined as:

$$d = h - h_p \tag{2.3}$$

where h is the head field without pumping, and h_p is the head field affected by pumping at steady state ($t \to \infty$).

The seepage velocity $\mathbf{v}$ [LT^{-1}] is related to the Darcy velocity $\mathbf{q}$ [LT^{-1}] through the porosity θ of the subsurface medium and is calculated as

$$\mathbf{v} = \frac{\mathbf{q}}{\theta} = \frac{-\mathbf{K}_h \nabla h}{\theta}. \tag{2.4}$$

For modeling contaminant transport with groundwater flow, I only consider conservative and non-reactive behavior. This allows to use the well-known advection-dispersion equation [e.g., *Fetter and Fetter*, 1999] to approximate the pore-scale transport processes on the REV scale by:

$$\frac{\partial c}{\partial t} + \nabla \cdot (\mathbf{v}c - \mathbf{D}\nabla c) = 0, \tag{2.5}$$

in which c [M L^{-3}] is the contaminant concentration and $\mathbf{D}$ [L T^{-1}] is the dispersion tensor. According to *Scheidegger* [1961], the entries D_{ij} of the dispersion tensor are defined as:

$$D_{ij} = \frac{v_i v_j}{\|\mathbf{v}\|}(\alpha_l - \alpha_t) + \delta_{ij}(\alpha_t \|\mathbf{v}\| + D_e), \tag{2.6}$$

where v_i is the i-th component of the seepage velocity, α_l [L] and α_t [L] are the longitudinal and transverse dispersivities, respectively, δ_{ij} is the Kronecker symbol and D_e [] is the effective porous medium diffusion coefficient. The required boundary conditions can again be defined by either Dirichlet or Neumann conditions:

$$\begin{aligned} c &= c_{di} \quad \text{on} \quad \Gamma_{di} \\ \mathbf{v}c - \mathbf{D}\nabla c &= \dot{\mathbf{m}}_{ne} \quad \text{on} \quad \Gamma_{ne}, \end{aligned} \tag{2.7}$$

where c_{di} is the fixed concentration value imposed at Dirichlet boundaries Γ_{di} and $\dot{\mathbf{m}}_{ne}$ is the specified contaminant mass flux density at Neumann boundaries Γ_{ne}. For steady-state flow and transport, the equations above can be simplified by removing all time derivatives.

2.1.2. Vadoze zone

In Chap. 4 I will apply different crop growth models in a test case as an example for conceptually different models. All models are implemented using the software package expert-N

[*Engel and Priesack*, 1993] and are based on the same numerical model for vertical transport of water, solute and heat in the unsaturated zone based on the Richards equation [*Richards*, 1931]:

$$\frac{\partial S}{\partial t} = \frac{\partial}{\partial z}\left[\mathbf{K}_h(S)\left(\frac{\partial h}{\partial z} + 1\right)\right] \tag{2.8}$$

where S [-] is the water content, z [L] is the elevation, $\mathbf{K}_h$ [LT^{-1}] is the hydraulic conductivity [-] and h [L] is the pressure head. The soil hydraulic functions are parametrized by the Mualem Van-Genuchten model [e.g., *Simunek et al.*, 2005]. Details about the crop growth models can be found in *Wöhling et al.* [2013].

2.2. Basic statistics

This section provides the required basic tools from statistics, probability theory and the Monte-Carlo approach for the stochastic characterization of random events.

2.2.1. Descriptive statistics

Methods from descriptive statistics are used to describe, interpret and analyze observed data sets using mathematical tools. Data sets are usually assumed to be observed outcomes of a random variable or of a stochastic process. The goal of descriptive statistic is to describe the general features of a random variable by summarizing characteristics (often called statistics of the data) by e.g., the average value and the average spread. Such simple characteristics can provide profound insights and reveal information that are otherwise hidden in the shear size of large data sets.

Univariate statistics

For a given data sample u, the most basic summary statistic is its arithmetic mean μ_u, which describes the central tendency of the data set by:

$$\mu_u = \frac{1}{n}\sum_{i=1}^{n} u_i, \tag{2.9}$$

where n is the total number of values u_i in the data set. The average spread of the data sample around the mean μ is quantified by the unbiased sample variance:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \mu)^2. \tag{2.10}$$

However, using squared differences changes the unit of variance and sometimes makes a physical interpretation difficult. The standard deviation σ is expressed in the same unit as u and therefore provides more meaningful values for physically-based quantities. It is the average deviation from the mean, which is the square root of the variance: STD $= \sqrt{\sigma^2}$.

Multivariate statistics

For two or more variables, descriptive statistics additionally provide measures to quantify and identify relationships or dependencies between variables. Let u_i, v_i with $i = 1, .., n$ be two data samples that are additionally described by their joint behavior and mutual dependency. One popular measure of linear dependency is the sample covariance between variables, which is defined as:

$$\mathrm{Cov}[u, v] = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \mu_u)(v_i - \mu_v) \tag{2.11}$$

A positive covariance $\mathrm{Cov}(X, Y)$ value states that the variable u shows similar behavior, which means that if u takes high values, v statistically tends to have a high value as well. An opposing behavior leads to negative covariance values. Linearly independent variables result in a covariance value of zero.

For linear systems, the covariance is a complete description of dependency between variables, but the covariance suffers from major deficits in nonlinear systems. The deficits of the covariance as a measure of dependency is of great importance within this thesis and the estimation of data impact in general. A more suitable measure of dependence is Mutual Information, which will be introduced in Sec. 2.4.2.

2.2.2. Probability theory

Probability theory is a branch of mathematics used to describe random events and random variables. The central idea is to describe events of one or several random variables by a probability functions. Probability distribution functions are more extensive and informative tools to describe (sets of) random variables than the scalar summary statistics from the previous section. In fact, while sample statistics are an incomplete description of samples, and samples from a random variable represent the true random variable again only an incomplete manner, only probability densities are exhaustive descriptions of a random variable.

Univariate case

The probability mass function (*PMF*) of a discrete random variable U defines the probability of each possible value u:

$$P_U(u) = P(U = u), \tag{2.12}$$

and is denoted by $P_U(u)$. *PMF*s describe the distribution of a discrete variable and can be illustrated using a histogram (see Fig. 2.1(a)). The histogram is the common estimation technique of distribution functions and is widely used. A cumulative mass function $F_U(u)$ is defined to describe the probability that a random variable is below any given value u

$$F_U(u) = P(U < u). \tag{2.13}$$

The *CMF* values take ranges between zero and unity, since the sum of the probabilities of all possible events is as well unity. Discrete variables often appear in the context of technical application, e.g., discrete grid indices or decision variables.

In contrast, parameters describing environmental quantities are rarely discrete, but rather continuous. Uncertain continuous quantities are described as continuous random variables. Let X be a continuous random variable in $\mathbb{R}$, then the number of possible values x of X is infinite. Therefore, a description by a PMF is inadequate, as the probability mass of a single event $P(X = x)$ is zero for all $x \in \mathbb{R}$ [*Weiss*, 2006]. However, it is possible to define a probability mass of an interval of values $P(X \leq x)$, which allows for the definition of the cumulative distribution function (*cdf*), denoted as F. The *cdf* describes the probability of X being below a value x

$$F_X(x) = P(X \leq x) \tag{2.14}$$

The *cdf* aggregates the probability over the range of X and therefore takes values in the interval between zero and unity (see Fig. 2.1(b)), just like *CMF*s do. A probability density function *pdf*, which is denoted by $p(x)$, can be defined as a nonnegative function defined in $\mathbb{R}$ such that

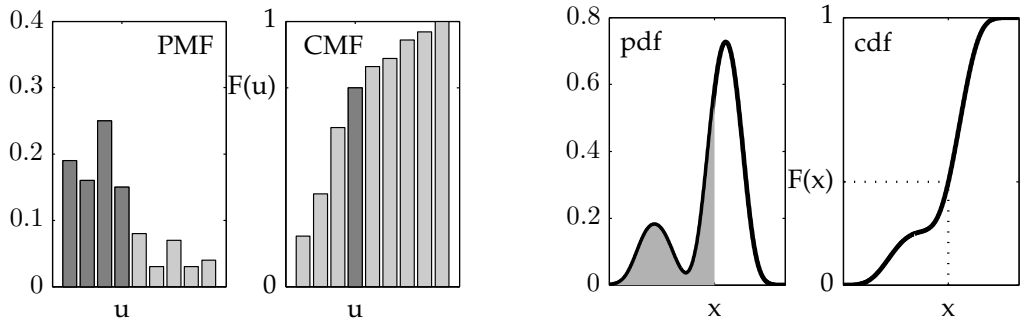
$$F_X(x) = \int_{-\infty}^{x} p(x)dx \tag{2.15}$$

Alternatively, the continuous *pdf* of x can be defined as the derivative of the *cdf*

$$p(x) = \frac{d}{dx} F_X(x) \tag{2.16}$$

for all values of x where $p(x)$ is defined. Note that, in contrast to probabilities, probability densities can exceed unity for narrow distributions. Similarly to PMF values, which sum up to unity, the basic property of any *pdf* is that its integral over the entire domain is equal to one:

$$\int_{-\infty}^{\infty} p(x)dx = \lim_{x \to \infty} F(x) = 1. \tag{2.17}$$



(a) Discrete random variable (b) Continuous random variable

Figure 2.1.: Examples for probability and cumulative distributions

The mean value μ_X of a random variable X with the *pdf* $p(x)$ is defined by the expected value $E[X]$, which is the integral of all values of the random variable weighted by their occurrence probability:

$$\mu_X = E_X[X] = \int_{-\infty}^{\infty} x\, p(x)\, dx \tag{2.18}$$

For a given finite sample x_i with $i = 1, .., n$, which is drawn from the theoretical distribution $p(X)$, the sample mean defined in Sec. 2.2.1 is an estimator for the theoretic mean of the distribution.

The variance of a random variable $V_X[X]$ is as well defined via its *pdf*:

$$V_X[X] = \int_X (x - \mu_X)^2 \; p(x) \; dx. \tag{2.19}$$

Multivariate case

The joint *pdf* of multiple random variables can be defined in a similar fashion via the joint *cdf*. Joint *pdf*s represent a complete description of all individual variables and their interdependencies. The marginal (individual) *pdf* of each variable can be obtained by integration of the joint *pdf* over the remaining variables. As an example, for two variables X, Y, integration over Y leads to the marginal *pdf* of X as

$$p(x) = \int_y p(x, y) dy. \tag{2.20}$$

The two variables X and Y are independent if the probability distribution of X does not depend on the value of Y. In this case, the joint probability density is the product of their marginal *pdf*s: $p(x, y) = p(x)p(y)$.

Conditional *pdf*s $p(x|y)$ and respectively $p(y|x)$ give the distribution of one variable for the case that the other variable is known. The conditional *pdf* shows directly how the observation of one variable changes the *pdf* of the other. The conditional *pdf*s of independent variables are equal to their marginal pdfs since they do not depend on the observed value.

The covariance between X, Y is defined as:

$$\begin{aligned} \text{Cov}[X, Y] &= E_X[((x - \mu_X)(y - \mu_Y)] \\ &= \int_Y \int_X (x - \mu_X)(y - \mu_Y) \; p(x, y) \; dx \; dy. \end{aligned} \tag{2.21}$$

and describes how two variables interact with each other. The magnitude of the covariance is dependent on the scale and therefore sometime hard to interpret. The correlation coefficient, which is a normalized and therefore dimensionless version of the covariance, is calculated as

$$\text{Corr}[X, Y] = \frac{\text{Cov}(X, Y)}{\sqrt{V[X] \; V[Y]}}. \tag{2.22}$$

Because $\text{Cov}(X, Y)$ and $\text{Corr}(X, Y)$ are linear measures of dependency, one can only conclude that if X and Y are independent then $\text{Cov}(X, Y) = \text{Corr}[X, Y] = 0$, but not vice versa. The correlation takes values within the interval $[-1 \; 1]$.

2.2.3. Density functions and their estimation

Previously, the *pdf* of some random variables was assumed to be known. In practical applications this is rarely the case. Instead, a distribution is often represented by a limited sample in practice. Parametric density estimation replaces the sample by a sufficiently close theoretic distribution function, by fitting the distribution parameters of a theoretical distribution to the data, e.g., using maximum likelihood methods. This allows to evaluate the *pdf* analytically later on. Many theoretical distribution functions are available, e.g., the Gaussian, log-normal, exponential or beta distribution, which are the most common ones. The most often used theoretical distribution function is the multivariate-Gaussian distribution that describes multiple correlated random variables that individually follow the Gaussian distribution. The multi-Gaussian joint-*pdf* of a k-dimensional random vector $\mathbf{x} = (x_1, .., x_k)$ is given by

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\mathrm{Cov}(\mathbf{x})|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathrm{Cov}(\mathbf{x})^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{2.23}$$

where $\boldsymbol{\mu}$ is the vector of mean values in k dimensions, and $\mathrm{Cov}(\mathbf{x})$ is the $k \times k$ covariance-matrix.

Parametric estimation is fast and usually requires only small sample sizes, since only a few parameters for the most suitable *pdf* out of a set of candidate distributions needs to be fitted. The decision for a certain theoretical function and therefore pre-defined shape lead to a simple estimation process by parameter fitting, however is not applicable for arbitrary distribution shapes.

In case that the sample resembles none of the available analytical distributions, it is possible to estimate arbitrary distributions with **non-parametric estimation** techniques. This class of estimators comprises a huge number of conceptually different approaches to approximate local density estimates. The class of non-parametric density estimators is solely data-driven, without any assumptions about the distribution shape. Therefore, non-parametric estimators are more flexible, but demand more data and are computationally more demanding. Especially high-dimensional density estimation is challenging and an ongoing research field. Selected non-parametric estimators are discussed in the next sections.

Histogram-based estimation

A sample $\mathbf{x}_i$ with $i = 1, .., n$ values of a random variable $\mathbf{X} \sim p(\mathbf{x})$ is given and represents a density that is to be estimated. Density estimation via histogram is the simples approach for density estimation, but is limited to low dimensions. The parameter space is partitioned into (usually equally sized) classes. The density in each class is estimated by the number of samples points within the class. However, in higher dimensions, classification approaches lead to bad stochastic approximations, since only few samples are located in each class, which is illustrated in the center of Fig. 2.2.3.

Kernel density estimation

Within the group of non-parametric estimators, I found kernel-based density estimators (KDE) to be the most suitable choice for arbitrary and possibly high-dimensional distributions. A good overview of existing kernel-based techniques is provided by *Scott* [2008]. Kernel estimators are based on the idea of expanding point information of a given data set (called source points $\mathbf{x}_s$ in the following) in space by using a kernel smoothing function K. The Kernel function depends on the distance between target points $\mathbf{x}_{t,j}$ and source points $\mathbf{x}_{s,i}$ and a defined bandwidth parameter. An illustration for a Kernel function applied to one source point is given in the left part of Fig. 2.2.3. The sample density at a target location $\mathbf{x}_{t,j}$ is estimated by the summed kernel values of all source points at the target point location and is calculated as follows:

$$\hat{p}_{\mathrm{KDE}}(\mathbf{x}_{t,j}) = \frac{1}{n}\sum_{i=1}^{n} w_{ij} K(\mathbf{x}_{t,j} - \mathbf{x}_{s,i}). \tag{2.24}$$

Above, K is the kernel function, which is often chosen to be Gaussian shaped. w_{ij} is an additional weighting matrix, which can be used to weight the source points in case a conditional *pdf* is estimated; otherwise the weighting is uniform. An example for a kernel-based *pdf* estimation is shown one the right side of Fig. 2.2.3. Comparing it to histogram-based estimation in the middle of Fig. 2.2.3 shows how the spatial expansion of the point information produces a smooth *pdf*, revealing more information than the noisy and oscillating histogram-based pdf.

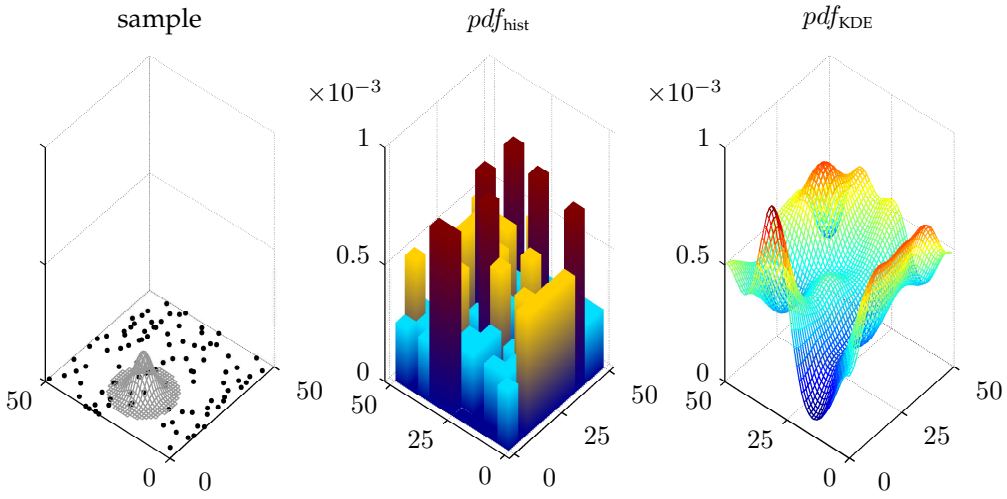


Figure 2.2.: Example for probability density estimation. The left plot shows a random sample and one kernel function for a single point. The center plot shows the results of a histogram-based density estimation for a 10×10 grid. The right side illustrates the *pdf* estimated using KDE techniques.

Using KDE techniques inherently leads to an additional smoothing of the *pdf*. To minimize this artificial smoothing, the kernel width is usually chosen as small as possible. Too small kernel widths, in contrast, do not allow to expand the point information sufficiently and thus lead

to noisy *pdf*s. For this reason, there exist an optimal kernel width that merely introduce the required (but no more than that) smoothing. *Silverman* [1986] suggested following relation that approximates the optimal kernel bandwidth as:

$$\hat{k}_{opt} = 1.06\hat{\sigma}_i n_p^{-1/(n_d+4)}, \tag{2.25}$$

where n_d is the dimensionality, n_p is the sample size, and $\hat{\sigma}_i$ is a sample-based estimate of the i-th marginal standard deviation.

However, related to random variables that describe potential measurement values subject to measurement errors, it is possible to exploit the artificial smoothing to account for the measurement error. Generally, an observed quantity contains unknown measurement error that is expressed through a variance σ_ϵ^2, which needs to be considered in the *pdf* estimation. Therefore, the measurement error σ_ϵ^2 is used to scale the width of kernel K. By doing so, this generates the *pdf* of the observable data including the error from the error-free simulated sample. For this special case, the obtained *pdf* in Eq. (2.24) is a non-smoothed representation of the *pdf* of observable values that include the random measurement error:

$$\hat{p}_{\text{KDE}}(\mathbf{x}) = p(\mathbf{x} + e_\mathbf{x}) \quad \text{with} \quad e_x \sim \mathcal{N}(0, \sigma_{\epsilon_x}^2). \tag{2.26}$$

For the summation of Gaussian kernels, three possible implementations are introduced in Appendix A. They differ on the implementation level as well as by the used computer resources.

2.2.4. Monte-Carlo methods

This section introduces the method of Monte-Carlo (MC), which is a brute-force and very general computation algorithm to produce samples of a random variable. It was named by Nicholas Metropolis and relies on repeated random experiments (simulated on computers). Conducting many repeated random experiments typically allows a numerical approximation of a random variable or process without analytical description. This approach is often used when closed-form solutions for*pdf*s are not available or deterministic algorithms are to expensive. The law of large numbers states that the Monte-Carlo approximation of any statistics converges to the true value for an infinite number of evaluations.

Monte-Carlo integration

A classical application of Monte-Carlo methods is the computation/estimation for high-dimensional integrals, when grid-based integration methods become too expensive. For example, the expected value of an arbitrary function $f(x)$ can be approximated as:

$$E_x[f(\mathbf{x})] = \int_x f(\mathbf{x})p(\mathbf{x})\, dx \approx \sum_{i=1}^{N} f(\mathbf{x}_i) \quad x_i \sim p(\mathbf{x}). \tag{2.27}$$

In contrast to grid-based methods, the convergence rate of MC is independent of the dimensionality of the integration. MC estimates generally converge at a rate of $O(1/\sqrt{(}N))$, where

N is the number of evaluations [*Lepage*, 1980]. Only the initial uncertainty depends on the dimensionality and the shape of the distribution. Therefore, MC methods are often applied to high-dimensional problems.

Monte-Carlo simulation

The idea of multiple random experiments within MC methods is especially valuable for stepping from deterministic models with known parameters to stochastic models that consider uncertain parameters. Arbitrary sources of uncertainty are modeled as random input variables. By drawing samples from the corresponding probability distribution, one obtains sets of input variables for the corresponding model. Each set of input variables is treated as one Monte-Carlo experiment. The general idea is to generate multiple, equally likely random experiments that resemble a sample of an unknown model output parameter. In post-processing, one can evaluate the statistics of that sample. Numerous Monte-Carlo methods exist, but mostly resemble similar patterns:

1. Generate multiple input parameter sets according to a defined *pdf* or sampling rule.
2. Apply a deterministic calculation or run a numerical model for each of the drawn parameter sets.
3. Aggregate and analyze the resulting sample of the relevant model output

In uncertain quantification, MC simulation is often used to estimate the uncertainty of a model output, specifically in the case of complex models where an analytical transfer from input uncertainty to prediction uncertainty is impossible. This can generally be solved by conditional Monte-Carlo simulations (see Sec. 2.3.5).

Since Monte-Carlo applies a series of (deterministic) experiments, it can be combined with any kind of simulation tool. It does not require any adaptations in the underlying software and therefore, is called non-intrusive. Such a non-intrusive stochastic framework can be applied to any system model and allows to extend deterministic models for stochastic simulations purposes. The general applicability is, however, paid by high computational costs, which limits the application to computationally expensive simulations models.

2.3. Bayesian modeling framework

This section will provide the basic methods to properly treat different types of uncertainty within numerical modeling of environmental systems.

2.3.1. Description of uncertainty

The lack of information is ubiquitous in environmental simulation problems, especially before any intensive site investigation efforts. Given the general inability to fully describe the under-

lying processes occurring in geosciences [*Oreskes et al.*, 1994], the selection of a single geostatistical, structural or conceptual model is often unjustifiable. Following this rationale, requires to incorporate various sources of uncertainty, which can be distinguished as:

- **Parametric uncertainty** is related to the model input and characterizes the insufficient information about the primary model parameters. The most important uncertain parameter for subsurface flow and transport models is hydraulic conductivity. It is a spatially correlated parameter, varying over several magnitudes, whose dependencies can be described using spatial correlation models. In model-based geostatistics, such models describe the spatial correlation using a set of structural parameters.
- **Structural parameter uncertainty** is a higher-order uncertainty description of incomplete knowledge about the structure of parametric uncertainty. This arises from the fact that the models used to describe parameter uncertainty are as well uncertain. For geostatistics, this leads to the assumption that the structural parameters of geostatistical models are uncertain as well. *Kitanidis* [1995] introduced the field of Bayesian geostatistics (see Sec. 2.3.3) to deal with such uncertain structures and to infer parameters under such conditions.
- **Conceptual model uncertainty** considers uncertainty in the appropriate choice of the system model. This includes different conceptual models (e.g. uncertain zonation, boundary/initial conditions or model forcing) as well as different numerical modeling implementations (e.g. finite-element-based transport or random-walk particle tracking), also including their numerical errors. Some parts of the literature list model input uncertainty as a different category. Bayesian model averaging (BMA) is one potential way to deal with model uncertainty and is introduced in Sec. 2.3.4.
- **Measurement uncertainty** accounts for the fact that observational data are inevitably subject to measurement errors. Numerous sources of errors exist and are hard to quantify even if the entire process of data acquisition is surveyed. Without specific information, the general approach is to characterize observation errors as additive or multiplicative error terms, which are often assumed to be Gaussian distributed and independent.

Any type of input uncertainty is propagated using the system model to delineate the resulting output or predictive uncertainty. As mentioned above, uncertain and therefore random variables are mathematically described by probability theory. Yet, there exist two fundamentally different views and interpretations of probability:

The **frequentist probability** defines probability based on occurrence of a possible outcome of a repeated experiment, which is also called relative frequency. This calculative approach is strictly objective and only based on the available data and does not allow for additional assumptions and soft knowledge. Without any data available, no frequentist probability can be evaluated and the approach is of limited usefulness.

The **subjectivist probability** is defined by the degree of belief about an occurrence probability. The most popular subjective probability is the **Bayesian probability**, which allows inference between probabilities extracted from data (e.g. samples from random experiments) and subjective prior probabilities, which result from possible expert knowledge.

The concept of Bayesian probability is well-suited for the description of uncertain models. It is especially useful for modeling based on little available data, where frequentist probabilities cannot be computed. It allows for a more flexible description of the state of knowledge including prior assumptions and expert knowledge, and therefore allows using all available types of information. In the Bayesian sense, the initial belief state is represented by the *prior* distribution. New information or evidence are incorporated via their *likelihood* to update the initial belief. The updated belief after inclusion of the new data is represented by the *posterior* distribution. The use of prior assumptions is always subject of discussion and needs to be justified adequately.

2.3.2. Bayes' theorem

The method to obtain the posterior distribution is called Bayesian inference, since the resulting posterior is inferred from the data by merging the new information about the system with the prior. The core of this inference is Bayes' theorem, which was named after Thomas Bayes and was further developed by Pierre-Simon Laplace [*Laplace*, 1820]. Let $p(\mathbf{s})$ be the prior distribution of the model parameters $\mathbf{s}$ and $\mathbf{y}$ be a set of new observation data. To infer the posterior (conditional) distribution of a parameter $p(\mathbf{s}|\mathbf{y})$, I apply Bayes' theorem:

$$p(\mathbf{s}|\mathbf{y}) = \frac{p(\mathbf{s},\mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{s})p(\mathbf{y}|\mathbf{s})}{p(\mathbf{y})}, \tag{2.28}$$

where $p(\mathbf{y}) = \int_{\mathbf{s}} p(\mathbf{s})p(\mathbf{y}|\mathbf{s})d\mathbf{s}$. A similar form of Bayes' theorem involves omitting the factor $p(\mathbf{y})$, which does not depend on $\mathbf{s}$ and can thus be considered a constant for a given data set $\mathbf{y}$. This leads to an unnomalized posterior distribution that is proportional to the actual posterior:

$$p(\mathbf{s}|\mathbf{y}) \propto p(\mathbf{s})p(\mathbf{y}|\mathbf{s}). \tag{2.29}$$

This is useful because $p(\mathbf{y})$ is often unknown and difficult to estimate. The fact that the integral of a *pdf* is unity can be used to transform and re-normalize the unnomalized posterior from Eq. (2.29). In fact, most conditional MC methods try to draw a sample from $p(\mathbf{s}|\mathbf{y})$ without the need of approximating $p(\mathbf{y})$.

2.3.3. Bayesian geostatistics

Let $\mathbf{s}$ be a $n_s \times 1$ vector of n_s distributed and heterogeneous subsurface parameters like hydraulic conductivity. Their spatial distribution is a result of natural processes like, e.g., sedimentation and therefore is not independent, but correlated between neighboring locations. The model-based geostatistical approach characterizes the spatial pattern of such parameters by a set of structural parameters $\boldsymbol{\theta}$. It uses space random functions (SRF) [e.g., *Journel and Huijbregts*, 1978; *Kitanidis*, 1997; *Rubin*, 2003] as models of the spatial structures. This approach can be used to simulate spatially dependent fields of uncertain hydraulic conductivity. The main features can be characterized by the spatial mean $\mu(\mathbf{x}) = E_{\mathbf{x}}[s(\mathbf{x})]$, the auto-variance $\sigma^2(\mathbf{x}) = E_{\mathbf{x}}[s(\mathbf{x}) - \mu(\mathbf{x})]$ and, most importantly, by the covariance matrix $C_{\mathbf{ss}} = \mathrm{Cov}(s(\mathbf{x}), s(\mathbf{x}')) = E_{\mathbf{x}}[(s(\mathbf{x}) - \mu(\mathbf{x}))(s(\mathbf{x}') - \mu(\mathbf{x}'))]$. Under the assumption of *second-order*

stationarity, mean and auto-variance become independent of the location $\mathbf{x}$ and the covariance function can be expressed as a functional of the separation distance $r = \|\mathbf{x} - \mathbf{x}'\|$ alone, independent of the individual locations.

Making assumptions about the correlation structure is powerful for two reasons: (1) it reduces the allowed parameter space by assigning a meaningful structural model and (2) the point information from observation data can be interpolated spatially. This is done by the well-known Kriging methodology [*Krige*, 1951]. Kriging is basically a geostatistical interpolation scheme that honors the underlying correlation structure. Kriging strongly relates to the *Best Linear Unbiased Estimator* (BLUE) [e.g., *Müller*, 2007] for intrinsic functions and solves the estimation problem of a given set of structural parameters and a given set of observations. For a detailed description of the different Kriging approaches, I refer to *Kitanidis* [1997]. In Sec. 2.3.5, I introduce linear co-Kriging as an example of linear Bayesian inference, which is, however, only one example from a broad spectrum of Kriging methods.

In conventional geostatistics, the structural model $\boldsymbol{\theta}$ is usually fitted using classical variogram analysis [e.g., *Matheron*, 1971] or maximum likelihood estimation [e.g., *Kitanidis*, 1996; *Schweppe*, 1973]. Although classical kriging does not rely on that assumption, the joint distribution of all values of $\mathbf{s}$ is often assumed to be multi-Gaussian: $p(\mathbf{s}) = \mathcal{N}(\mu(\mathbf{s}), C(r))$.

The most commonly chosen covariance model is the Gaussian model, which is defined as

$$C_{Gauss}(r) = \sigma^2 \exp\left(-\frac{r^2}{\lambda^2}\right). \tag{2.30}$$

The Gaussian covariance results a second-order stationary function with the highest possible entropy in the correlation structure which is and often associated with the goal of making the least amount of assumptions about a distributions shape. However, other differently shaped covariance models are available such as, e.g., exponential, nugget or the power covariance model. They differ, among others, in the asymptotic behavior of long-rang correlation and in the smoothness of the described fields.

The Gaussian covariance, the multi-Gaussian model and two-point statistics in general receive criticism for being too simple and inflexible to realistically describe real parameter distributions in the subsurface. Notably for transport simulations, multi-Gaussian fields produce connected path lines around the mean value and lead to unrealistic transport times, which is discussed in *Zinn and Harvey* [2003]. There exists an extensive research field of multi-point statistics that aims to overcome the simplifications of Gaussian approaches using, e.g., rank order statistics [e.g., *Journel and Deutsch*, 1997], copulas [*Bárdossy*, 2006] or multi-point Geostatistics based on training algorithms [e.g., *Mariethoz et al.*, 2010].

Besides the special criticism about two-point statistics, any specific selection of a correlation structure is often hard to justify with limited available data. Therefore, *Bayesian geostatistics* [e.g., *Diggle and Ribeiro*, 2007; *Kitanidis*, 1997] are often used as an extension of geostatistics that allows for uncertain structural parameters. Instead of choosing one set of structural model parameters $\boldsymbol{\theta}$, the Bayesian approach allows for multiple or even distributions of structural parameters modeled by a distribution $p(\boldsymbol{\theta})$. The parameter set $\boldsymbol{\theta}$ thus becomes a set of random variables of its own. The so-called Bayesian distribution is calculated by marginalization over the structural parameters $p(\boldsymbol{\theta})$ [*Kitanidis*, 1986]:

$$\tilde{p}(\mathbf{s}) = \int_{\boldsymbol{\theta}} p(\mathbf{s}|\boldsymbol{\theta}) \ \mathrm{d}\,\boldsymbol{\theta}, \tag{2.31}$$

which does not dependent on one given set of $\boldsymbol{\theta}$, but on the given distribution $p(\boldsymbol{\theta})$. Bayesian geostatistics are closely related to Bayesian model averaging, which is introduced in the next section.

2.3.4. Bayesian model averaging

Facing high and omnipresent uncertainty for sparsely investigated systems makes it difficult to setup a single justifiable model. Any corresponding assumptions to restrict oneself to a unique model selection are hard to defend prior to deeper site investigation. Instead, a selection or an entire spectrum of models may be adequate, and considering several competing models can help to reduce the subjectivity of the selection. Thus, one needs to admit a selection of model alternatives and should weight them according to their credibility include subjectivity. The simulation task is performed for all considered models, and posterior model credibilities are assigned after comparison with available data. This procedure is called *Bayesian model averaging* (BMA) [e.g., *Hoeting et al.*, 1999; *Neuman*, 2003].

Let k be an indicator variable for identification of a set of n_k different discrete models. BMA evaluates the posterior distribution of a model prediction z, by assimilating the observation data $\mathbf{y}$ in all considered models M_k and by subsequently averaging over all models:

$$p(z|\mathbf{y}) = \sum_{k=1}^{n_k} p(z|M_k, \mathbf{y}) p(M_k|\mathbf{y}). \tag{2.32}$$

This allows to infer averaged posterior statistics such as mean, variance or entropy. Bayes theorem is also used to infer the posterior model credibility by:

$$p(M_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k)p(M_k)}{\sum_{l=1}^{n_k} p(\mathbf{y}|M_k)p(M_k)} \tag{2.33}$$

after *Hoeting et al.* [1999]. The idea is that such an averaged response of a suite of models is superior to each individual one.

A recent advancement of BMA shifts the problem of discrete model choice to a parametric model selection. The novelty is that it allows for a continuous spectrum of model alternatives [*Feyen*, 2003; *Nowak et al.*, 2009; *Murakami et al.*, 2010], which are parameterized as yet an other random variable. For example, the authors parameterized the choice among different covariance models via the Matérn family of covariance functions [*Matérn*, 1986] that allows different correlation lengths as well as different shapes. It is defined as

$$\begin{aligned} C(l) &= \frac{\sigma}{2^{\kappa-1}\Gamma(\kappa)} (2\sqrt{\kappa} l)^{\kappa} B_{\kappa}(2\sqrt{\kappa} l) \\ l &= \sqrt{\sum_{i}^{n_d} \left(\frac{\Delta x_i}{\lambda_i} \right)^2}, \end{aligned} \tag{2.34}$$

where $\Gamma(\cdot)$ is the Gamma function and n_d is the number of spatial dimensions. $B_\kappa(\cdot)$ is the modified Bessel function of the third kind [*Abramowitz and Stegun*, 1972]. The additional shape parameter κ controls the shape of the covariance function and includes as well the conventional covariance models, e.g., the exponential model with $\kappa = 0.5$ and the Gaussian model with $\kappa = \infty$. The benefits of the Matérn family have been discussed extensively by, e.g., *Handcock and Stein* [1993] and *Diggle and Ribeiro* [2002]. The relevance of the Matérn family within Bayesian Geostatistics was pointed out by *Nowak et al.* [2009]. Continuous BMA is an elegant extension to BMA, especially in MC frameworks, which handles the model choice parameter κ as just another random variable.

Fig. 2.3 illustrates how two discrete covariance models (Gausian and exponential) are included in the general Matén family.

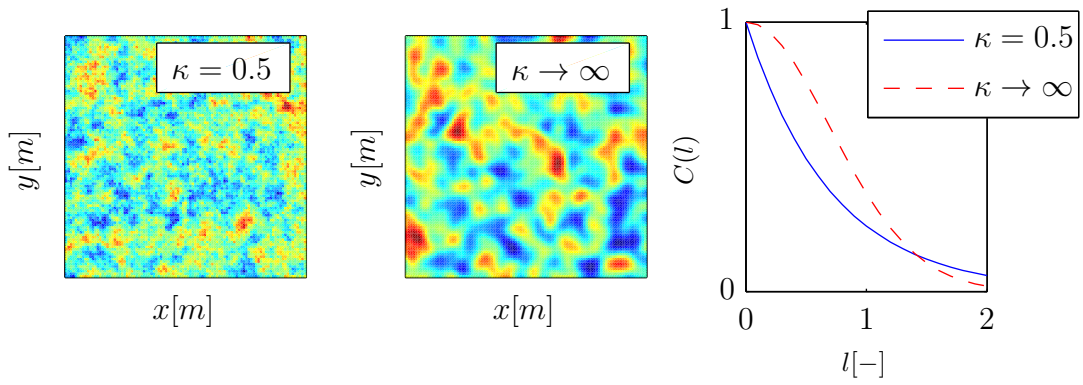


Figure 2.3.: Example random fields of hydraulic conductivity with a correlation length of $\lambda_x = \lambda_y = 10m$ and different Matern parameter κ. The left plot shows a exponential covariance function for $\kappa = 0.5$ and center plot shows second a Gaussian shape with $\kappa = 50$. Both functions are plotted over the distance in correlation lengths on the right plot.

We simply include κ in the vector of uncertain structural parameters $\boldsymbol{\theta}$. The Bayesian distribution of the parameter field is then obtained by marginalizing the classical geostatistical description over all unknown meta-parameters $\boldsymbol{\theta}$:

$$\tilde{p}(\mathbf{s}) = \int_{(\boldsymbol{\theta})} p(\mathbf{s}|\boldsymbol{\theta}) \ \mathrm{d}(\boldsymbol{\theta}) \tag{2.35}$$

Bayesian inference used to infer posterior distributions of the model outcome or of the structural parameters is an essential part of BMA. The Bayesian geostatistical approach is equivalent to the continuous BMA (see Eqs. (2.32)-(2.33)) and is included in the next section, which describes in detain the entire inference process under consideration of all types of uncertainty.

2.3.5. Bayesian inference, analytical solutions and bootstrap filter

Bayesian inference is used to infer posterior distributions of parameters and predictions from the prior distribution and the available observation data. In contrast to parameter calibration or maximum likelihood estimates, Bayesian posterior distributions are combinations of prior belief and observation, and they quantify the remaining uncertainty after considering the data. The next section introduces in detail the general Bayesian update within an uncertain model setup.

Prior Bayesian distribution

Modeling the prior model belief state $p(\mathbf{s}, \boldsymbol{\xi}, k, \boldsymbol{\theta})$ is often done by using the available data to define an adequate individual prior distribution for the different meta parameters $\boldsymbol{\xi}, k, \boldsymbol{\theta}$. As mentioned, $p(\boldsymbol{\theta})$ reflects the prior structural model distribution, $p(\boldsymbol{\xi})$ the uncertain boundary and initial conditions and $p(k)$ the conceptual model belief state. The geostatistical approach allows to model the distribution of parameters $p(\mathbf{s}|\boldsymbol{\theta})$ given the structural parameters $\boldsymbol{\theta}$. The combined prior belief state is then represented by the joint distribution:

$$p(\mathbf{s}, \boldsymbol{\xi}, k, \boldsymbol{\theta}) = p(\mathbf{s}|\boldsymbol{\theta})p(\boldsymbol{\xi})p(k)p(\boldsymbol{\theta}), \tag{2.36}$$

with $p(\mathbf{s}) = \int_{\boldsymbol{\theta}} p(\mathbf{s}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\,\boldsymbol{\theta}$. It is of course possible to model the meta parameters as dependent parameters with a joint distribution, in case that the available informations substantiate this assumption.

Bayesian update

Assuming there is a $n_y \times 1$ vector of measurement values $\mathbf{y}$ related to the parameters $(\mathbf{s}, \boldsymbol{\xi}, k, \boldsymbol{\theta})$ through a model $\mathbf{y} = \mathbf{f}_y(\mathbf{s}, \boldsymbol{\xi}, k, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_y$, where $\boldsymbol{\varepsilon}_y$ follows some distribution that accounts for the (typically white-noise) measurement error and sometimes also model structure error. The model response $\mathbf{f}_y$ is the simulated true value according to the model. For a given realization and all parameters $(\mathbf{s}, \boldsymbol{\xi}, k, \boldsymbol{\theta})$, the residuals between measurements $\mathbf{y}$ and corresponding model response $\mathbf{f}_y(\mathbf{s}, \boldsymbol{\xi}, k, \boldsymbol{\theta})$ are modeled as measurement and model error $\epsilon_y \sim \mathcal{N}(0, R_\epsilon)$. This allows deriving the likelihood of any given parameter set. For instance, assuming a Gaussian error distribution for $\boldsymbol{\varepsilon}_y$ (similar to *Feyen* [2003], or *Christensen* [2004]) yields the likelihood function

$$L(\mathbf{y}|\mathbf{s}, \boldsymbol{\xi}, k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}_y(\mathbf{s}, \boldsymbol{\xi}, k, \boldsymbol{\theta}), \mathbf{R}_{\boldsymbol{\varepsilon}}), \tag{2.37}$$

where $\mathbf{f}_y(\mathbf{s}, \boldsymbol{\xi}, k)$ is the mean and $\mathbf{R}_{\boldsymbol{\varepsilon}}$ is the error covariance matrix. In particular, assuming measurement errors and model structure errors to be Gaussian and independent is common use in many fields of science and engineering, including data assimilation [e.g., *Evensen*, 2007]).

The conditional distribution of $\mathbf{s}$ given $\mathbf{y}$ and all meta parameters $\boldsymbol{\xi}, k, \boldsymbol{\theta}$ is determined by Bayes theorem:

$$p\left(\mathbf{s}|\boldsymbol{\xi}, k, \boldsymbol{\theta}, \mathbf{y}\right) \propto L(\mathbf{y}|\mathbf{s}, \boldsymbol{\xi}, k, \boldsymbol{\theta})\; p(\mathbf{s}|\boldsymbol{\xi}, k, \boldsymbol{\theta}). \tag{2.38}$$

Of course, assuming known meta parameters $\boldsymbol{\xi}, k, \boldsymbol{\theta}$ cannot be justified prior to extensive data collection (see Sec. 2.3.4). The conditional marginal distribution of $\mathbf{s}$ given only $\mathbf{y}$, called a Bayesian distribution [*Kitanidis*, 1986], is given as:

$$\tilde{p}(\mathbf{s}|\mathbf{y}) \propto \int\limits_{(\boldsymbol{\xi},k,\boldsymbol{\theta})} p(\mathbf{s}|\boldsymbol{\xi},k,\boldsymbol{\theta},\mathbf{y})\; p(\boldsymbol{\xi},k,\boldsymbol{\theta}|\mathbf{y})\,\mathrm{d}\,(\boldsymbol{\xi},k,\boldsymbol{\theta})\,, \tag{2.39}$$

where the tilde denotes the Bayesian probability. Note that the entire distribution $p(\mathbf{s},\boldsymbol{\xi},k,\boldsymbol{\theta})$ has been jointly conditioned on $\mathbf{y}$.

In general, the task of a model is to predict a dependent $n_z \times 1$ vector of n_z predictions $\mathbf{z}$ related to $\mathbf{s}$, $\boldsymbol{\xi}$, $\boldsymbol{\theta}$ and k via a model $\mathbf{z} = \mathbf{f}_z(\mathbf{s},\boldsymbol{\xi},k,\boldsymbol{\theta})$. Typically, $\mathbf{z}$ does not have an additional independent stochastic component other than the ones discussed above. One could, however, easily replace the set of noise-free predictions $\mathbf{z}$ by noisy predictions $\mathbf{z}'$ using yet another error $\boldsymbol{\varepsilon}_z$ with an appropriate distribution. In a hydro(geo)logical context, $\mathbf{z}$ might be a water level connected with a free-surface flow model or a contaminant concentration, and $\mathbf{f}_z$ might be a flow and transport equation such as the ones in Sec. 2.1. The conditional prediction $\tilde{p}(\mathbf{z}|\mathbf{y})$ then becomes

$$\tilde{p}(\mathbf{z}|\mathbf{y}) \propto \int\limits_{(\mathbf{s},\boldsymbol{\xi},k,\boldsymbol{\theta})} p(\mathbf{z}|\mathbf{s},\boldsymbol{\xi},k)\; p(\mathbf{s}|\boldsymbol{\xi},k,\boldsymbol{\theta},\mathbf{y})\; p(\boldsymbol{\xi},k,\boldsymbol{\theta}|\mathbf{y})\;\,\mathrm{d}\,(\mathbf{s},\boldsymbol{\xi},k,\boldsymbol{\theta})\,, \tag{2.40}$$

where $p(\mathbf{z}|\mathbf{s},\boldsymbol{\xi},k)$ is the raw distribution that reflects $\mathbf{z} = \mathbf{f}_z(\mathbf{s},\boldsymbol{\xi},k)$.

Within the focus of BMA, the posterior distribution of meta parameters conditional on the available data is calculated as:

$$p(\boldsymbol{\theta},\boldsymbol{\xi},k|\mathbf{y}) = \frac{L(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\xi},k)p(\boldsymbol{\theta},\boldsymbol{\xi},k)}{p(\mathbf{y})} \tag{2.41}$$

with

$$p(\mathbf{y}) = \int\limits_{(\boldsymbol{\theta},\boldsymbol{\xi},k)} p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\xi},k)p(\boldsymbol{\theta},\boldsymbol{\xi},k)\,\mathrm{d}(\boldsymbol{\theta},\boldsymbol{\xi},k). \tag{2.42}$$

Analytical solutions

For given meta parameters $\boldsymbol{\theta},\boldsymbol{\xi},k$ and linear model relations between data and parameters, there exist analytical solutions for Bayesian updating, such as e.g., Kriging [*Krige*, 1951]. The key approach is based on linear error propagation and the corresponding $n_s \times n_y$ sensitivity matrix (or Jacobian) $\mathbf{H}$, which gives the sensitivity of all data values $\mathbf{y}$ with respect to all parameter values $\mathbf{s}$ as $\mathbf{y} = \mathbf{f}_y(\mathbf{s}) = \mathbf{H}\mathbf{s} + \epsilon_y$. The entries of the sensitivity matrix are given as:

$$H_{ij} = \frac{\partial \mathbf{f}_{y_i}(\mathbf{s})}{\partial s_j}. \tag{2.43}$$

One needs to distinguish between two types of measurements: Direct measurements are related to the model input parameters (e.g. conductivity, porosity) itself, whereas indirect measurements are measurements of model output quantities (e.g. hydraulic head, drawdown, concentrations). For direct measurements, the sensitivity matrix $\mathbf{H}$ is easy to compute and is unity for the observed parameters and zero for all parameters. For indirect measurements, the sensitivity matrix $\mathbf{H}$ can be derived through sensitivity analysis methods such as adjoint state sensitivities [*Sykes et al.*, 1985] or by straightforward numerical differentiation as in PEST [*Dougherty and Marryott*, 1991]. Linear error propagation yields the covariance matrix between observations and parameters $\mathbf{Q}_{sy}$ and the auto-covariance of the measurements $\mathbf{Q}_{yy}$:

$$\begin{aligned} \mathbf{Q_{sy}} &= \mathbf{Q_{ss}}\mathbf{H}^T \\ \mathbf{Q_{yy}} &= \mathbf{H}\mathbf{Q_{ss}}\mathbf{H}^T + \mathbf{R}_\epsilon \end{aligned} \tag{2.44}$$

where $\mathbf{R}_\epsilon$ is the error covariance of the measurements and $\mathbf{Q_{ss}}$ is the covariance of the parameter field $\mathbf{s}$. Using covariances that describe the linear dependencies allows to evaluate the expected value of the posterior parameter distribution $\hat{\mathbf{s}}$ as:

$$\hat{\mathbf{s}} = \bar{\mathbf{s}} + \mathbf{Q_{sy}}\mathbf{Q_{yy}}^{-1}(\bar{\mathbf{y}} - \mathbf{H}(\mathbf{s})), \tag{2.45}$$

where $\bar{\mathbf{s}} = E[\mathbf{s}]$ is the known mean and $\bar{\mathbf{y}} = H\bar{\mathbf{s}}$ is the linear response. This equation is identical with the simple Kriging equations *Krige* [1951]. More complex Kriging principles is available for, e.g., unknown and uncertain mean or parameter fields including a spatial trend.

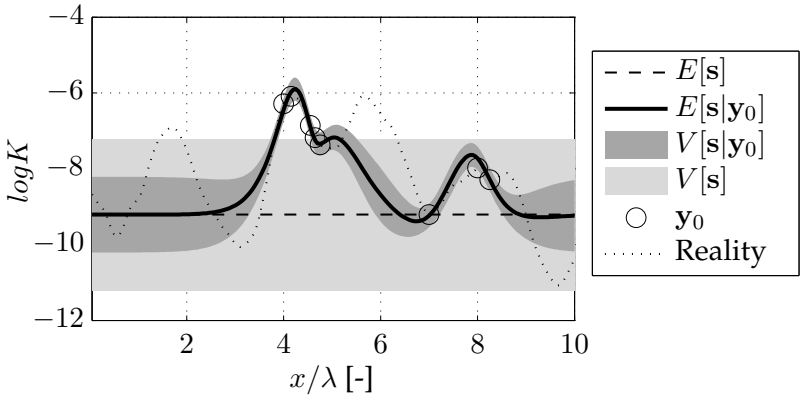


Figure 2.4.: Example for Kriging of a one-dimensional random field. The shaded gray areas illustrate the prior and the posterior field variance around the mean. Measurements $\mathbf{y}_0$ are indicates by the circles, which are taken from the synthetic reality (dotted line).

The advantage of linear models is that the posterior covariance is independent of the actual measured value an can be estimated via:

$$\mathbf{Q_{ss|y}} = \mathbf{Q_{ss}} - \mathbf{Q_{sy}}\mathbf{Q_{yy}}^{-1}\mathbf{Q_{ys}}, \tag{2.46}$$

An example of linear updating by Kriging is given in Fig. 2.4. The conditional mean $E(\mathbf{s}(\mathbf{x}))|\mathbf{y})$ depends on the value of $\mathbf{y}$, but the posterior field variance $\text{Var}(s(\mathbf{x})|\mathbf{y})$ does not. The conditional prediction $\hat{\mathbf{z}}$ is often computed via MC by generating parameter fields with the properties given in Eqs. (2.45)-(2.46), respectively, via conditional simulation. If the model $f_z(\mathbf{s})$ is also sufficiently linear, then Eqs. (2.44)-(2.46) can alternatively be applied directly to the prediction z for direct linear error propagation, yielding the posterior prediction $\hat{\mathbf{z}} = \mathbf{H}_z\hat{\hat{\mathbf{s}}}$ and $\mathbf{Q}_{\mathbf{zz}|\mathbf{y}} = \mathbf{H}_z\mathbf{Q}_{\mathbf{ss}|\mathbf{y}}\mathbf{H}_z^T$.

The Ensemble Kalman filter (EnKF) [*Evensen*, 1994] is a method that calculates covariances required for Eqs. (2.45)-(2.46) based on an available ensemble of MC realizations. Initially, the Kalman filter and the related EnKF were designed for real-time updating in transient forecasting problems when new data is available. Due to their popularity, they have been extended for parameter estimation [e.g., *Chen and Zhang*, 2006; *Nowak*, 2009a; *Schöninger et al.*, 2012] without a time-dependent step.

For nonlinear cases, the equations provided above may be used in an approximated sense, based on a linearization of the model. Then, however, the conditioning and estimated posterior covariances will not be exact. Nonlinear dependencies can be approximated by e.g., iterative updating, because the linear sensitivity matrix $\mathbf{H}$ is only a local tangent to the model. Successful implementations of linearized updating to weakly nonlinear problems can be found in, e.g., *Kitanidis* [1995]; *Yeh et al.* [1996]; *Nowak* [2009a]. Sequential updating schemes [e.g., *Vargas-Guzmán and Yeh*, 2002] are applied first on direct data about parameter and secondly the indirect data, as the former usually pose linear problems and only the latter require iteration. However, strong nonlinearities can pose challenges to such iterative approaches and prevent a convergence to proper solutions or make the iteration costly. The major problem is still that linear(ized) updating rules applied to nonlinear problems are only an approximations, and this approximations are more or less crude, depending on the degree of nonlinearity.

Bootstrap filter

The bootstrap filer (BF) [*Gordon et al.*, 1993], also known as particle filter or condensation, is a nonlinear and ensemble-based updating scheme. It is basically a sequential importance sampling algorithm, which is often combined with resampling strategies at the end. The basic idea is to first draw equally likely realizations from the prior distribution and then apply a realization-wise weighting that is proportional to their likelihood to match the observation data. The resulting weighted ensemble represents the posterior distribution.

Brute-force filtering on observation data captures any degree of nonlinearity and requires no assumptions of the distribution shapes. In the limit of infinite sample size, the BF becomes an accurate brute-force implementation of Bayesian updating. In contrast to analytical solutions, it allows for uncertain structural parameters $\boldsymbol{\theta}, \boldsymbol{\xi}, k$, which can therefore be included as additional random variables. Hence, BMA can be implicitly included in the BF.

For n realizations of $(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi}, k)$ independently drawn from $p(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi}, k)$ and a (hypothetically given) data set $\mathbf{y}_0$, the BF would evaluate an $n \times 1$ normalized weight vector $\mathbf{w}$ with

$$w_i = \frac{L((\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi}, k)_i|\mathbf{y}_0)}{\sum_{j=1}^{n} L((\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi}, k)_j|\mathbf{y}_0)}, \tag{2.47}$$

where $L()$ denotes the likelihood function of the parameters for the given data.

The advantage of filtering techniques to capture nonlinear dependencies is paid by tremendous computational costs, especially for high-dimensional filtering of large data sets. The ensemble of generated prior realizations needs to be sufficiently large, otherwise only few fitting realizations receive substantial weights. This is known as filter degeneration [e.g., *Liu*, 2008; *Snyder et al.*, 2008; *Van Leeuwen*, 2009], indicating that the weighted sample is not reliable for statistical estimation anymore. In spite of a possibly large weighted ensemble, only few realizations contribute to the posterior. This is specifically problematic when normalizing of the weights leads to artificially high weights of realizations, whose likelihood is still very small, but they receive large weights, merely because they are the best available ones within an overall poor set. In addition, realizations with low weights remain in the ensemble and require the same computational effort as large-weight realizations, but do not contribute significantly to the results. These limitations, also called the *curse of dimensionality*, have been the scope of many studies in the past [e.g., *Liu*, 2008; *Snyder et al.*, 2008; *Van Leeuwen*, 2009].

One common way to avoid filter degeneracy is resampling [e.g. *Snyder et al.*, 2008], where realizations with low weights are discarded and additional realizations are generated in order to stabilize the posterior estimation. A suitable measure to survey the quality of a weighted sample is the effective sample size (ESS) [*Liu*, 2008], which is an empirical estimation of the number of particles with substantial weights:

$$ESS = \left(\sum_{i=1}^{n} (w_i)^2 \right)^{-1}. \tag{2.48}$$

The ESS assumes values in the interval of $[0, n]$ and estimates how many samples effectively contribute to the posterior ensemble.

Applying th weights to Eq. (2.47) to all realizations results in a weighted posterior ensemble that can be used to extract any desired posterior statistics. Weighted averaging of n prediction realizations $z_i = \mathbf{f}_z((\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi}, k)_i)$, $i = 1...n$, would yield the posterior expectation

$$E_{z|\mathbf{y}_0}[z] \approx \frac{1}{v_1} \sum_{i=1}^{n} z_i \, w_i, \tag{2.49}$$

with $v_1 = \sum_{i=1}^{n} w_i$. The variance of a weighted sample is computed as:

$$V_{\mathbf{z}|\mathbf{y}_0}[z] \approx \frac{v_1}{v_1^2 - v_2} \left[\sum_{i=1}^{n} z_i^2 \, w_i - \left(\sum_{i=1}^{n} z_i \, w_i \right)^2 \right], \tag{2.50}$$

with $v_1 = \sum_{i=1}^{n} w_i$ and $v_2 = \sum_{i=1}^{n} w_i^2$. $E_a[b]$ is the expected value of b over the distribution of a and $V_a[b] = E_a[b^2] - E_a[b]^2$ is the respective variance. Here both quantities are computed in a most efficient way after [*Weiss*, 2006, p. 355]. The corresponding correction factor in

Eq. (2.50) resembles the well-known factor $\frac{1}{n-1}$ for the non-weighted sample variance. This is an unbiased estimator of the population variance even for small effective sample sizes.

I will discuss other statistics describing uncertainty in Sec. 2.4, which can also be estimated from weighted ensembles.

2.4. Information theory

Information theory is the part of mathematics and computer science that describes and quantifies information. It roots back to the work of Claude E. Shannon in the field of signal processing. It provides fundamental measures of information and interdependency - e.g. *Entropy*, *Relative Entropy* and *Mutual Information*, which are functionals of probability distribution functions. In the field of experimental design, the *Fisher information* or *Information matrix* is widely used in linear or linearized statistical models. These terms will be defined in the following.

2.4.1. Fisher information

Fisher information $\mathcal{F}$ [e.g., *Edgeworth*, 1908] defines a measure of how much information a set of random variables $\mathbf{Y}$ possesses about a set of unknown or uncertain model parameters $\mathbf{X}$:

$$\mathcal{F}(\mathbf{X}) = p(\mathbf{Y}|\mathbf{X}) \left(\frac{\partial}{\partial \mathbf{X}} \log p(\mathbf{Y}|\mathbf{X}) \right) \left(\frac{\partial}{\partial \mathbf{X}} \log p(\mathbf{Y}|\mathbf{X}) \right)^T . \tag{2.51}$$

Fisher information is a common measure for information in experimental design. For linear (or linearized) models, the inverse of the posterior covariance matrix is the Fisher information matrix. Therefore, the goal of minimizing the posterior variance corresponds to maximizing the Fisher information. Maximizing or minimizing of matrix-valued quantities requires first to map these matrices onto a scalar quantity (such as the trace or determinant), which is a non-unique choice. This non-uniqueness leads to the alphabetic optimality criteria [*Box*, 1982].

Working with Fisher information is intrinsically connected to the variance as a measure of uncertainty. Apart from linearized models, the Cramer-Rao inequality proves the inverse Fisher matrix to be the lower bound of the conditional covariance matrix of the parameters. In other words, the Fisher information is a measure of the theoretical estimation precision limit, to which the model parameters $\mathbf{X}$ can be estimated. This means that the Fisher information is only a true information measure for linear models .

2.4.2. Entropy

In Information theory, the basic measure for the uncertainty of random variables is *Shannon entropy*. Similar to thermodynamics, the entropy of a set of continuous variables $\mathbf{x}$ and a set of

discrete variables $\mathbf{X}$, respectively, is a measure of disorder associated with the variables, and is defined as:

$$\begin{aligned} H(\mathbf{X}) &= -\sum_{\mathbf{X}} P(\mathbf{X}) \log P(\mathbf{X}) \qquad \text{for discrete } \mathbf{X} \\ h(\mathbf{x}) &= -\int_{\mathbf{x}} p(\mathbf{x}) \log \mathbf{x} \qquad \text{for continuous } \mathbf{x}, \end{aligned} \tag{2.52}$$

where H is the discrete entropy and h is the differential entropy. The entropy of a binary variable is a special case of discrete entropy, in which $\mathbf{X}$ takes only values of zero or unity.

It is a common convention that $0 \log 0 = 0$, which is necessary as $\log 0$ is not defined. More uncertain and therefore more broadly distributed variables with smaller density values lead to higher entropies, whereas narrow distributions lead to small entropies. Differences in the definition of discrete and differential entropy lead to some important aspects that need to be considered. Firstly, when approaching perfect certainty about a random number, the discrete entropy converges to zero, which is also its lower bound, whereas the differential entropy can take negative values. Secondly, the *pdf* of a continuous random number does not always exist (e.g., if the *cdf* is not differentiable), and in these cases a differential entropy does not exist either.

Relative entropy h_{rel}, also known as the Kullback-Leibler divergence [*Kullback and Leibler*, 1951], can be used to measure the difference between the conditional distribution $p(\mathbf{x}|\mathbf{y})$ and the prior distribution $p(\mathbf{x})$. It quantifies the change in the distribution of $\mathbf{x}$ that is caused by conditioning $\mathbf{x}$ on a given set of observation data values $\mathbf{y}_0$ (see Fig. 2.5). It is defined as:

$$h_{rel}\left(\frac{\mathbf{x}|\mathbf{y}}{\mathbf{x}}\right) = \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}_0) \log \frac{p(\mathbf{x}|\mathbf{y}_0)}{p(\mathbf{x})} d\mathbf{x}. \tag{2.53}$$

The data $\mathbf{y}_0$ is affecting the investigated *pdf*s of $\mathbf{x}$, and the relative entropy measures their difference and, therefore, is related to the information value $\mathbf{y}_0$.

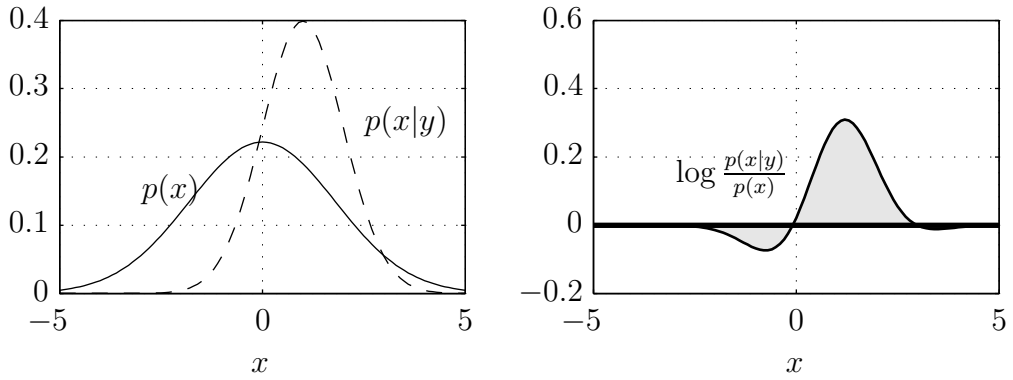


Figure 2.5.: Example of a prior and posterior distribution and the integrand of the relative entropy or Kullback-Leibler distance.

However, before $\mathbf{y}_0$ is actually measured within data acquisition, its actual value is unknown, and thus is the case for any experimental design. Therefore, the relative entropy needs to be evaluated in an expected sense. This is done by integrating h_{rel} over the (possibly high-dimensional) distribution of potential data values $p(\mathbf{y})$.

2.4.3. Mutual Information

The expected value of relative entropy is also known as *Mutual Information* and is defined as:

$$\mathrm{MI}(\mathbf{x},\mathbf{y}) = E_{\mathbf{y}}\left[h_{rel}\left(\frac{\mathbf{x}|\mathbf{y}}{\mathbf{x}}\right)\right] = \int_{\mathbf{y}} p(\mathbf{y}) \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log\left[\frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})}\right] \, d\mathbf{x}\, d\mathbf{y}. \tag{2.54}$$

Mutual information is a general measure of the expected information which one random variable provides about another. It requires no assumption about the random distributions and the order of dependency. I therefore consider MI as the most accurate measure of expected information content. Other measures of information, like Fisher information or linear sensitivities, can be linked to MI under certain assumptions and simplifications.

MI can also be linked to conditional entropy, which is defined as:

$$h(\mathbf{x}|\mathbf{y}) = \int_{\mathbf{y}} p(\mathbf{y}) \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log\left[p(\mathbf{x}|\mathbf{y})\right] \, d\mathbf{x}\, d\mathbf{y}. \tag{2.55}$$

Thus, MI is alternatively defined as the difference between prior entropy $h(\mathbf{x})$ and conditional entropy $h(\mathbf{x}|\mathbf{y})$:

$$\mathrm{MI}(\mathbf{x},\mathbf{y}) = h(\mathbf{x}) - h(\mathbf{x}|\mathbf{y}). \tag{2.56}$$

This is graphically illustrated by the Venn diagram (see Fig. 2.6) from set theory that shows the connection of entropy, relative entropy and mutual information.

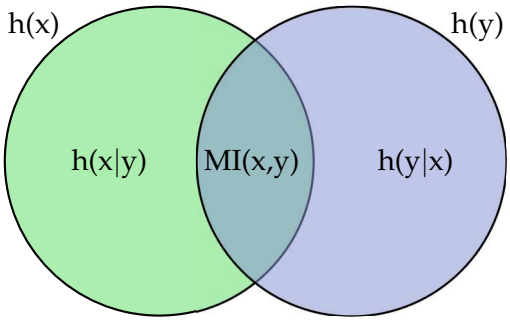


Figure 2.6.: Venn diagrams can be constructed for low numbers of random variables and illustrate the interaction between entropy, conditional entropy and mutual information.

Variance as information measure

The variance of a random variable is only an approximation of the uncertainty, but actually measures the spread of a distribution. Only for parametric distributions with fixed shapes,

variance can be guaranteed to be monotonically related to entropy. Therefore, one can state that applying variance as uncertainty measure implicitly assumes a fixed multi-Gaussian distribution shape. In such cases, the reduction of the variance within an inference from prior to posterior is a suitable measure of information for the used measurement data. For distribution shapes that are only close to the Gaussian, but otherwise free in their shape, the variance is only an approximation. For arbitrary shapes (e.g., multi-modal distributions), the variance can lose any connection to information at all.

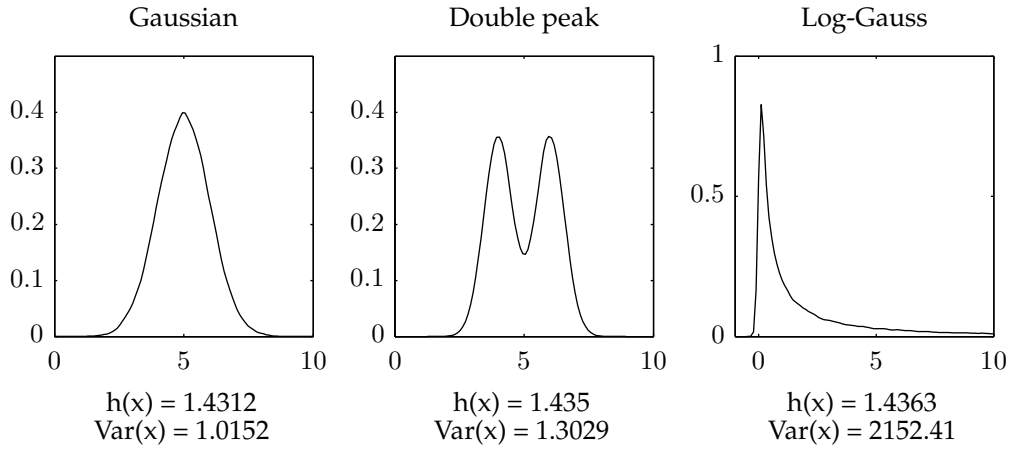


Figure 2.7.: Example for the differences and deficits of variance as an uncertainty measure for different distribution types.

Fig. 2.7 shows a comparison between variance and entropy for different shaped distributions. All distributions are chosen to result in the same entropy. For the bimodal distribution, the variance is slightly increased as the values (compared to the Gaussian distribution with the same entropy) of the distribution move away from the mean. For the log-normal distribution on the right side, the variance is increased by three orders of magnitude. This shows the deficits of the variance to describe the uncertainty of a variable and its limitations to serve as a substitute for true information measures.

3. Optimal design of data acquisition

This chapter continues to introduce basic methods in the context of *design of experiments* (DoE) in general and specifically for designing data acquisition. The methods of this chapter represent statistical analysis in the context of optimal design and are based on many previously introduced methods from the last chapter. Sec. 3.1 defines the general DoE problem, its mathematical description and the parametrization of design decisions within data acquisition campaigns. Of course, efficient search algorithms are indeed a key factor in effective DoE methods, but are not the scope of this thesis. Therefore, I rely on state-of-the-art algorithms that are introduced in Sec. 3.2. Sec. 3.3 introduces the concept of data impact and discusses the deficits of linear data impact estimates in anticipation of the nonlinear estimates in Sec. 4.

3.1. General definition

The planing and design of data acquisition originated from linear and nonlinear regression models [e.g., *Box*, 1982; *Pukelsheim*, 2006], was later extended towards the calibration of computer models [e.g., *Gates and Kisiel*, 1974] and finally treated as a DoE problem [e.g., *Federov and Hackl*, 1997]. In hydrogeology, this concept was extended to *Geostatistical optimal design* [e.g., *Cirpka et al.*, 2004; *Herrera and Pinder*, 2005; *Müller*, 2007] to infer spatially correlated parameters under the assumption of a known structural model (see Sec. 2.3.3). The field of *Bayesian experimental design* (as reviewed in *Chaloner and Verdinelli* [1995]) considers multiple competing models that can not be excluded at the current state of knowledge. It accounts for uncertain model choice and provides the expected utility, averaged over all models (see BMA in Sec. 2.3.4). Its extension in Geostatistics, namely *Bayesian geostatistical optimal design* [e.g., *Diggle and Lophaven*, 2006; *Nowak et al.*, 2009] avoids the undefendable assumption of one selected structural model, by allowing a spectrum of concurrent structural models. Another important development is the transition to task-driven DoE, which focuses on reducing the uncertainty in the model output related to one specific engineering task (see Sec. 3.3.2), rather than focusing on a generic reduction of parameter uncertainty.

Design and design space

The initial step, prior to any DoE endeavor, is a proper problem parameterization of the decision problem. Let $\mathbf{d}$ be a $n_{\mathbf{d}} \times 1$ design vector of $n_{\mathbf{d}}$ independent decision variables that specify the available choices concerning the execution of the data acquisition. This may comprise of, e.g., the number of measurements, their locations, the physical quantity to be measured and

other system forcing conditions, which are related to the measurement acquisition. Each individual choice needs to be parameterized and the allowed range of values needs to be determined. All allowed combinations of values in $\mathbf{d}$ span the $n_{\mathbf{d}}$-dimensional *design space* $\mathbf{D}$ of allowable designs. Constrains that affect the allowable design space $\mathbf{D}$ might, e.g., consider any financial and technical restrictions and apparently weak designs, as far as assessable beforehand.

Utility function

The next step in DoE is a proper description of the goal of the experimental design and how it can be measured. This goal is defined as a mathematical function and is generally called *utility function* or *objective function* $\Phi(\mathbf{d})$ [*Fishburn*, 1970; *Pukelsheim*, 2006], which is only dependent on the design $\mathbf{d}$. Specifically for data acquisition, the design utility is the anticipated value of information arising from the data, collected according to the design.

In many applications it is useful to express the data impact in monetary terms, so-called data-worth [e.g., *James and Gorelick*, 1994; *Feyen and Gorelick*, 2005]. For example *Liu et al.* [2012] developed a framework to assess the value of data within a subsurface remediation problem in monetary terms to allow balancing with the costs of data collection. Within this thesis, data impact is treated in a more general and not necessarily monetary way. If necessary, monetary aspect could be considered similar as in *Liu et al.* [2012]. Therefore, I define data utility via statistical and information-theoretic measures from Sec. 2.4.

Properly defining the utility function is most important within DoE. Since the latter optimization will blindly exploit any (even unreasonable) possibility to maximize the functional output, any shortcoming in the formulation of the utility will mislead the resulting design. Thus, the utility function must only depend on relevant decisions variables concerning the goals of data acquisition, and evaluated the degree of goal attainment (e.g., gained information). Choosing its numerical implementation or approximation also determines which degree of statistical dependencies is considered and the stochastic approximation quality. It will be the core of this entire thesis to introduce, develop and assess different concepts for utility functions, which all differ with respect to accuracy, evaluation speed and robustness against uncertain prior assumptions.

For certain utility functions is useful to distinguish between design utility $\Phi(\mathbf{d})$, which only depends on $\mathbf{d}$. In contrast data utility $\phi(\mathbf{d}, \mathbf{y}_0)$ depend on the actual measurement values $\mathbf{y}_0$ and can only be evaluated to a particular given one. Furthermore, all utility functions are dependent on the current model state, which is normally the model state prior to data collection. Starting from Chap. 6, the underlying model is changing and the dependency on the assumed model will be additionally indicated as a functional dependency of the utility function $\Phi(\mathbf{d}, \mathbf{M})$.

Optimization problem

The task of any DoE method is to determine which design $\mathbf{d}$ inside the design space $\mathbf{D}$ performs best related to the defined goal. The utility function allows for comparison of different designs

by a corresponding quality measure. Therefore, the DoE problem is formulated as a classical mathematical optimization problem. The optimization goal is to find that very design, which is superior to all other allowable designs. The so called *optimal design* $\mathbf{d}_{opt}$ is indeed defined as the design that maximizes the utility function:

$$\mathbf{d}_{opt} = \arg\max_{\mathbf{d}\in\mathbf{D}}[\Phi\{\mathbf{d}\}]. \tag{3.1}$$

The optimal design, of course, depends on the allowable design space $\mathbf{D}$ and on the formulation of the utility function, i.e., on goal settings and restrictions in the data acquisition campaign.

Solving this classical optimization problem requires specialized algorithms, which are introduced in the next section. For complex optimizations problems in high-dimensional design spaces, it is extremely hard to identify the one global optimal design. Therefore, it is often sufficient to evaluate so-called near-optimal designs which provide a utility value close to the global optimum. Such designs can be evaluated at much lower computational costs and provide an adequate cost-benefit ratio.

3.2. Optimization algorithms

The search within the design space $\mathbf{D}$ is one of the main challenges in optimization problems. In most cases, the space is extremely large, making it impossible to systematically evaluate all possible combinations. Just to illustrate this challenging task, I compute the number of possible designs of one used test case from Sec. 4.5.1 within this thesis. In this example, I consider a spatial grid of 260×160 locations, where measurements may be collected. At each location, two types of available data types my be collected, which leads to $n = 8320$ possible single measurements. The total number of combinations for the placement of $k = 10$ measurements, without two placing at the same place, is computed as:

$$N = \binom{n}{k} = \frac{n!}{(n-k)!k!}\ , \tag{3.2}$$

which leaves us with approximately 4.3×10^{42} combinations. The search in such huge spaces created an entirely new research field for the development of smart and efficient search algorithms that aim to find the optimum by only evaluating a small fraction of the design space. The '*No free lunch theorem*' [*Wolpert and Macready*, 1997] states, that there is not one algorithm that performs best in any class of search problems. In contrast, algorithms that exploit particular properties of the specific problem, like e.g., convex-shaped or smooth utility functions, always perform better than unspecific and generalized algorithms.

Available options are local (sequential) search algorithms such as Greedy Search (GS) [*Christakos*, 1992, p. 411]. This reduce the number of evaluated designs by independently optimizing one design variable at a time. However, such localized search algorithms fail to find the true global optimum. An extension to attenuate this deficit is called sequential exchange algorithm (SE) and will be introduced in Sec. 3.2.1.

Global search algorithms, such like particle swarm optimization [e.g., *Kennedy et al.*, 2001], Monte-Carlo-based optimization [e.g., *Haario et al.*, 2001] or simulated annealing [e.g., *Cerny*,

1985; *Kirkpatrick et al.*, 1983], use meta-heuristics to improve their search methods. The class of evolutionary algorithms like Genetic Algorithms (GA) [e.g., *Goldberg*, 1989; *Reed et al.*, 2000] imitate the biological evolution process of gene pools. The driving force of evolution is that the survival (and thus reproduction) of a design within a large population is dependent on its fitness. The absence of any individual superior algorithm leads to the class of adaptive multi-algorithm strategies such as AMALGAM-SO [*Vrugt et al.*, 2008] or genetic algorithm with multiple adaptive evolution schemes.

Recently, the focus of optimization in DoE shifted to multi-objective optimization, where different and possibly competing utility functions are optimized without being aggregated to a combined utility function. The result of such multi-objective problems are higher-dimensional pareto fronts of so-called non-dominated solutions, which allows the user to choose the best trade-off between concurring objectives. Some commonly used algorithms in this area are the CMA-ES [*Hansen and Kern*, 2004], the ϵ-hBOA [*Kollat et al.*, 2008] and the recent BORG algorithm [*Hadka and Reed*, 2013; *Reed et al.*, 2013]. BORG is an evolutionary multi-algorithm framework that offers different evolution concepts, which are adaptively chosen with respect to their effectiveness in solving the particular problem.

Although they are among the key factors for an efficient optimization, search algorithms are not the scope of this study. Instead, I rely on state-of-the art algorithms. A sequential exchange algorithm is introduced in Sec. 3.2.1 and a self-coded implementation of a GA is introduced in Sec. 3.2.2. Since my findings will be mostly independent of the used optimization algorithms, the algorithms are only introduced briefly. For more details, I refer to the respective cited literature.

3.2.1. Sequential exchange

Sequential exchange [e.g., *Christakos*, 1992] is an improved greedy search (GS) algorithm [e.g., *Cormen et al.*, 1990, p. 329]. Greedy algorithms simplify high-dimensional search problems by following a simple locally (dimension-vice) optimization heuristic. Thus, multi-point-designs are optimized by sequentially optimizing one sample point after the other. For the placement of the n-th measurement location, each possible location in the entire domain is evaluated, leaving all previously optimized $n-1$ design points fixed. This results in so-called data impact maps [*Nowak et al.*, 2009, 2012; *Leube et al.*, 2012] for each featured measurement type, which are very suitable to illustrate and discuss the optimization results in a direct and physically understandable fashion. However, greedy-like search algorithms are unable to find the true optimum in most problem classes, because each dimension is optimized independently [e.g., *Cormen et al.*, 1990].

After the multi-point design is optimized by greedy search, the sequential exchange algorithm restarts the loop with the first dimension and re-evaluates the all alternative decision choices, now in combination with the other fixed design choices. This can reveal a superior initial design solution that replaces the old one. After fixing the newly found optimal decision the same procedure is applied to the next dimension. This iterative loop is repeated until an entire cycle did not lead to any changes, which is defined as convergence of the sequential exchange

algorithm. The re-iteration reduces the dependency of previously taken decisions and leads to better changes for finding solutions closer to global optimum.

3.2.2. Genetic algorithm

Genetic Algorithms [e.g., *Goldberg*, 1989] (GA) mimic biological evolution. The technical equivalent of evolution replaces the survival fitness function by the utility function, and therefore the population evolves towards the global optimum of the function. GAs are global search algorithms to find sufficiently good near-optimal designs in high-dimensional design spaces. I use a self-implemented GA for global optimization of high-dimensional (e.g., multi-measurement) designs. It contains an elite survival scheme to preserve the superior design within the population. A tournament selection scheme [e.g., *Brindle*, 1981] is used for selecting the design to generate a new generation. One part of the next generation is generated via different crossover schemes:

- Sequence crossover,
- Multi-point crossover [*Jong and Spears*, 1991],
- Bitwise crossover,
- Differential evolution [*Storn and Price*, 1997] and
- Simulated binary crossover [*Deb and Agrawal*, 1994].

Which crossover functions are used is chosen adaptively based on their success rate, which is monitored throughout the optimization. A second fraction of the new generation is evolved by mutation schemes that alter the spatial coordinates of the measurement locations randomly:

- Local mutation,
- Global mutation and
- Poly mutation [*Deb and Agrawal*, 1994],

again selected adaptively with respect to their success rates. A small remaining fraction consists of newly generated, fully random designs to diversify the population. The success rate of the different evolution schemes are estimated based on an archive of the currently best obtained solutions. Other generally used parameters for the adjustment of a genetic algorithm are given in Tab. 3.1, together with the chosen values in this thesis.

3.3. Measures of data impact

This section will discuss various existing measures of data impact. It will also show that the complexity of estimating the impact of data increases with the complexity of the involved numerical model. Indirect observation quantities lead to the very challenging field of inverse modeling. Especially for data that trigger nonlinear inverse problems, accurately estimating their data impact poses large computationally challenges.

Generational parameters	
Population size	100
Number of generations	500
Archive size	300
Tournament size	2
Elite count	1
Evolution parameters	
Crossover fraction	0.40
Mutation fraction	0.40
Randomly fraction	0.20

Table 3.1.: General parameter setup of the genetic algorithm within this thesis.

A variety of alternative terminologies exists for data impact assessment (DIA). This includes, among others, the Value of Information Analysis (VOIA) [e.g., *Howard*, 1966; *Yokota and Thompson*, 2004], the data worth analysis, which originates in water resources from *James and Gorelick* [1994], and Bayesian decision analysis mentioned by *Feyen and Gorelick* [2005]. The VOIA methodology provides a useful set of relevant information measures related to the information of data. Three basic quantities of information value are:

1. The *Expected Value of Perfect Information* (EVPI) is the upper bound of data impact, when perfect and extensive observation data reduce all uncertainties to zero. Hence, it is also an estimate of the loss, (e.g., in prediction quality or decision confidence) caused by uncertainty.

2. Secondly, after data collection, the *Value of Measured Information* (VMI) can be evaluated, which is the actual data impact. This is a posterior analysis in the Bayesian sense and can only be done once the data values are known.

3. At the planning stage of data acquisition, any future data is yet unknown and only the *Expected Value of Measurement Information* (EVMI) can be estimated. This is typically the utility function $\Phi(\mathbf{d})$ in DoE of data acquisition.

The estimation process and the chosen measure of data impact $\Phi(\mathbf{d})$ are the core of this thesis, because they are the key factors for accuracy, computational costs and robustness. The general goal of data acquisition is to identify model parameters $\tilde{p}(\mathbf{s}|\mathbf{y}(\mathbf{d}))$ or reduce the uncertainty in the model prediction $\tilde{p}(\mathbf{z}|\mathbf{y}(\mathbf{d}))$. Extensive lists of utility measures are discussed, e.g., in *Federov and Hackl* [1997]; *Chaloner and Verdinelli* [1995]; *Müller* [2007]; *Nowak* [2009b]. It remains for the modeler to choose which measure and which approximation is most adequate for the given optimization problem. This choice includes, among other things, the order of statistical dependency between data and prediction goals that is considered. In a Bayesian sense, the estimation of data impact is based on Bayesian inference or related numerical approximations (see Sec. 2.3.5). Finally, one needs to decide which measure of information is adequate, for example, measures based on variance or entropy. In the end, most information measures can be related to either the Fisher or Mutual Information from information theory (see Sec. 2.4).

3.3.1. Linear estimates and their limitations

This section introduces the most common linear data impact estimates that can be found in literature. For models and data that pose linear or weakly nonlinear inverse problems, it suffices to measure the dependency between observation data, model parameters and eventually model predictions with linear or linearized methods. This is closely related to taking the assumption of multi-Gaussian dependency.

In general, the variance within model parameters is often used as a measure for the knowledge about a model. Therefore, the variance reduction by data can serve as a potential approximation of data impact. The basis of linear(ized) approaches to estimate variance reduction and work with a measure of sensitivity between potential observation data and inferred parameters or the relevant model output, respectively. The approaches can be distinguished by their methods to evaluate the dependencies between observation data and model parameters:

(1) One widely used category is the first-order second-moment (FOSM). The key advantage of FOSM is the straightforward and computationally efficient first-order propagation of mean and variance from parameters and data to predictions via sensitivity matrices H [e.g., *Kitanidis*, 1995] (see Sec. 2.3.5). The included sensitivity analysis is executed locally and measured by the partial derivatives of model output of interest with respect to all model parameters at a predefined value of the parameters, most commonly the mean value. By doing so, the covariance of the measurements can be approximated as :

$$\mathbf{Q_{yy}} = \mathbf{H}\mathbf{Q_{ss}}\mathbf{H}^T \tag{3.3}$$

Available codes like UCODE [*Poeter and Hill*, 1998] and PEST [*Doherty*, 2002] gained popularity due to their easy implementation and perform the numerical differentiation within the parameter space. This requires a large number of forward simulation for large parameter spaces. Successful applications with regard to DoE of data acquisition are presented by, e.g., *Kunstmann et al.* [2002] and *Cirpka and Nowak* [2004]. Based on covariances, FOSM is limited to linear and weakly nonlinear problems [e.g., *Schweppe*, 1973]. Within geostatistical applications, the size of the involved auto-covariance matrix of parameters can pose difficulties for large-grid models unless using spectral methods [e.g., *Nowak et al.*, 2003; *Fritz et al.*, 2009].

(2) In contrast to cumbersome forward differentiation codes, *Adjoint state* sensitivity analysis [e.g., *Sykes et al.*, 1985; *Sun and Yeh*, 1990; *Sun*, 1994; *Cirpka and Kitanidis*, 2001] uses an inverse modeling approach to provides local sensitivities as well. Because the perturbations for evaluating the sensitivities are conducted in the data space, it is particularly efficient for cases of small data spaces compared to the parameter space. Adjoint states are subject to the same physics as the original state, but with a reversal of causality or time. Such adjoint methods are most efficient to obtain sensitivity matrices for various sources of uncertainty. However, this requires intrusion into simulation codes and is often impossible for commercial software tools. *Cirpka and Nowak* [2004] used first-order second-moment-based methods combined with adjoint-state sensitivity analysis and Fast Fourier transform-based methods for geostatistical inversion [*Nowak et al.*, 2003; *Nowak and Litvinenko*, 2013].

(3) In contrast, a global sensitivity analysis provides the sensitivity over the entire range of relevant parameter values, which is important in cases that the system is possibly weakly

nonlinear. A strictly non-intrusive and straightforward method which avoids the handling of vast auto-covariance matrices is the *Ensemble Kalman Filter* (EnKF) [e.g., *Evensen*, 2007]. The EnKF has been extended towards geostatistical inversion and parameter estimation [e.g., *Zhang et al.*, 2005; *Nowak*, 2009a; *Schöninger et al.*, 2012]. Successful applications in OD can be found in, e.g., *Herrera and Pinder* [2005] and *Zhang et al.* [2005]. The main restriction of the EnKF is that it is optimal only for multi-Gaussian dependence among all involved variables, therefore *Schöninger et al.* [2012] transformed non-Gaussian shaped variables to attain optimal results.

The linear sensitivity matrix $\mathbf{H}$ allows updating any dependent variable by linear Bayesian inference [*Rouhani*, 1985] as described in Sec. 2.3.5 and hence generate a posterior state. All linearized approaches exploit the fact that posterior uncertainty is independent of the actual (yet unknown) future measurement values within the linear view [e.g., *Deutsch and Journel*, 1997], and thus depends only on the design. Therefore, the expected data impact can be estimated as the reduction in variance within the parameters while only knowing the design specifications:

$$\Phi(\mathbf{d})_D = V(s) - V(s|\mathbf{y}). \tag{3.4}$$

Eq. (3.4) holds for a single scalar parameter. For multiple parameters $\mathbf{s}$ it is required to compute a scalar measure from the reduction of a covariance. Two often applied possibilities to obtain scalar measures are:

$$\begin{aligned} \Phi(\mathbf{d})_C &= \quad \mathrm{tr}\left[\mathbf{Q}_{\mathbf{ss}} - \mathbf{Q}_{\mathbf{ss}|\mathbf{y}(\mathbf{d})}\right] \quad ; \quad \text{C-criterion}, \\ \Phi(\mathbf{d})_D &= \det\left[\mathbf{Q}_{\mathbf{ss}} - \mathbf{Q}_{\mathbf{ss}|\mathbf{y}(\mathbf{d})}\right] \quad ; \quad \text{D-criterion}, \end{aligned} \tag{3.5}$$

where tr is the matrix trace and det is the determinant of the matrix. Other general scalar DoE criteria can be found in *Box* [1982], which have been transferred to the geostatistical context by *Nowak* [2009b].

3.3.2. Task-driven data impact formulation

In DoE, task-driven approaches [e.g., *Ben-Zvi et al.*, 1988; *Nowak*, 2008; *Neuman and Ye*, 2009; *Nowak et al.*, 2009] are widespread. These doe not primarily focus on the identification of basic model parameters $\mathbf{s}$, but on the uncertainty reduction of a task-related model output z, e.g., the c-criterion of OD [e.g., *Sykes et al.*, 1985; *LaVenue et al.*, 1995; *Cirpka and Nowak*, 2004] or environmental performance metrics [e.g., *de Barros et al.*, 2012]

By focusing on the model output quantity, only those input parameters which are relevant for the model output are considered valuable for inference by new data [e.g., *Nowak et al.*, 2009]. A prominent example is the design of a remediation process based on an uncertain numerical model [e.g., *Bear and Sun*, 1998; *Copty and Findikakis*, 2000].

The task-driven formulation of linear data impact estimation requires a second sensitivity matrix $\mathbf{H}_z$ of the model prediction with respect to the model parameters, which is defined as:

$$\mathbf{H}_z = \partial z / \partial \mathbf{s}, \tag{3.6}$$

with this linearization, one can provide the estimated prediction variance by $\sigma_z^2 = \mathbf{H}_z \mathbf{Q}_{zz} \mathbf{H}_z^T$. Alternatively, the EnKF can directly compute the corresponding variance If one defines a reduction in prediction variance as overall goal, one obtains a task-driven DoE criterion.

$$\Phi(\mathbf{d})_C = \sigma_z^2 - \mathbf{H}_z^T Q_{\mathbf{ss}|\mathbf{y}} \mathbf{H}_z. \tag{3.7}$$

$\mathbf{H}_z$ mainly effects which parameters are valuable for calibration in order to obtain better estimates of z. Such linear information estimates are only accurate for truly linear DoE problems [e.g., *Müller*, 2007]. In other cases, designs optimized under linear information measures are called locally optimal designs [*Federov and Hackl*, 1997, p.100].

3.3.3. Towards nonlinear data impact estimation

For a flexible and generally applicable optimization framework, the restriction to linear or weakly nonlinear systems is not acceptable, especially in complex environmental systems. The need to consider various sources of uncertainty, different model structures, several physical processes and considering indirect measurement types introduces several components of nonlinearity. I will emphasize the most important sources that are relevant in the scope of subsurface systems:

1. The physical *forward operators*, relating the input parameters to the model outcome and to possible observable quantities, may include nonlinear equations, such as in multi-phase or unsaturated flow (see Richards equation in Sec. 2.1.2).

2. *Inverse operators* are generally nonlinear, even if the forward operators are linear partial different equations (see *Tarantola* [2005]). This is the case, e.g., for the groundwater flow equation and for the advection-dispersion equation (see Sec. 2.1.1).

3. *Structural* and *model uncertainty* as in Bayesian geostatistics [*Kitanidis*, 1995] introduce nonlinear dependencies between possible data and the model parameters, even if the particular individual models are linear. This nonlinearity requires special inversion techniques as discussed in, e.g., *Diggle and Lophaven* [2006]; *Nowak et al.* [2009].

4. *Higher-order geostatistics* have been used to overcome the traditional multi-Gaussian description of natural correlation structures [e.g., *Gomez-Hernandez and Wen*, 1998]. Such approaches as copula-based approaches [*Bárdossy and Li*, 2008], multi-point, multi-model or transition probability models introduce higher-order multi-variate dependencies between parameters. In these cases, linear conditioning approaches have been proven inadequate [*Kerrou et al.*, 2008].

5. *Task-driven* formulations in DoE introduce yet another possible source of nonlinearity. If the sub-surface model is included in a modular model framework, its output might be input for the next subsequent model, which may include other nonlinear operators. Examples are monetaristation frameworks as in *James and Gorelick* [1994], subsequent human health risk assessments [e.g., *de Barros et al.*, 2011] or prediction of exceedance probabilities of critical values [e.g., *Nowak et al.*, 2008; *Neuman and Ye*, 2009].

The list above includes many new modeling frameworks that aim to overcome previously used simplifications in sub-surface processes, dependencies and modeling. With these more sophisticated methods finding their way in practical applications, linear inversion methods and data impact measures pose a severe restriction for most design tasks. For nonlinear systems, multi-Gaussian or linear approaches become inadequate. Therefore, consistent and nonlinear estimation methods for data impact are urgently required.

I will focus on particular options for nonlinear data impact estimation in Sec. 4. From these, I chose the most flexible and efficient approaches to develop an accurate and nonlinear estimation framework in Sec. 4.2.

4. Nonlinear data impact assessment

The previous chapter pointed out the need for nonlinear schemes to assess data impact for subsurface systems. The current chapter is equivalent to **Step (I)** of my overall approach, i.e., to set up an accurate, flexible and therefore widely applicable reference framework for nonlinear data impact. Such a reference framework will allow in later chapters to benchmark and evaluate other, possibly approximated, estimates in terms of evaluation speed and estimation quality. Parts of this chapter have been published in a similar form in *Leube et al.* [2012], introducing this reference method, which was named *Preposterior data impact assessor* (PreDIA). It furthermore presents an application example from the field of contaminant hydrogeology. The second application presented here features an example from crop plant modeling (Sec. 4.6) and was published in *Wöhling et al.* [2013].

The new framework introduced in this chapter extends nonlinear Bayesian inference to estimate nonlinear data impact by marginalizing data impact over the distribution of yet unknown data values. The approach in Sec. 4.2 is a strictly formal information processing scheme and free of linearizations. It is an ensemble-based scheme that works with arbitrary simulation tools, provides full flexibility concerning measurement types (linear, nonlinear, direct, indirect) and allows for any desired task-driven formulation (see Sec. 3.3.2). Furthermore, it can account for arbitrary sources of uncertainty (see Sec. 2.3.1) via Bayesian geostatistics (see Sec. 2.3.3) and via Bayesian model averaging (see Sec. 2.3.4).

Existing methods fail to provide these crucial advantages all at once. The PreDIA method presented here achieves these advantages at relatively high computational costs (see Sec. 4.4.3). I demonstrate the applicability and advantages over conventional linearized methods in a synthetic example of subsurface transport in Sec. 4.5. The numerical example shows that informative data can be invisible for linearized methods that confuse zero correlation with statistical independence. Finally, I extend the example to specifically highlight the consideration of conceptual model uncertainty (see Sec. 4.5.3).

This is followed by a second application within the modeling of flow processes in the unsaturated zone combined with crop growth models in Sec. 4.6. The application features four competing models and shows how to re-evaluate a sampling campaign using a fully nonlinear data impact measure. This measures the potential of the investigated data to discriminate different conceptual models best.

4.1. Introduction

Linear versus nonlinear data impact analysis: The major difference of nonlinear data impact estimation to linear estimates is that the impact is dependent on the future measured data val-

ues. Since these are still unknown before data collection, only the expected data impact or EVMI (see Sec. 3.3) can be estimated. As shown by *James and Gorelick* [1994], *Feyen and Gorelick* [2005] and *Diggle and Lophaven* [2006], the EVMI or expected data impact can be evaluated by, e.g., Monte-Carlo-based averaging of posterior states over the anticipated statistical distributions of the still unknown data values. Averaging over synthetic posterior states led to the general term of *preposterior analysis* [e.g., *Ben-Zvi et al.*, 1988; *Freeze et al.*, 1992; *Feyen and Gorelick*, 2005] and is the core of nonlinear data impact analyses [*Ammar et al.*, 2011]. Several studies [e.g., *James and Gorelick*, 1994; *Diggle and Lophaven*, 2006; *Feyen and Gorelick*, 2005] used brute-force Monte-Carlo approaches to account for nonlinear dependencies between observable quantities, model parameters and model prediction, which are often present in complex environmental design problems (see details in Chap. 3.3.3). Within this analysis, there exist multiple posterior states, each one anticipating a set of measurement values and having the same probability to be in accordance with the actual future posterior state. The problem gains further complexity when considering model structural uncertainty, which leads, amongst other possible approaches, to Bayesian model averaging [e.g., *Hoeting et al.*, 1999].

Possible schemes for Bayesian inference: As a matter of principle, any sufficiently accurate conditioning method or Bayesian inference scheme can be employed in nonlinear DoE. This opens the path to all Monte-Carlo (MC) based methods. Most of the available MC-based techniques, i.e., the Pilot Point method (PP) [e.g., *RamaRao et al.*, 1995], the method of anchored distributions (MAD) [e.g., *Murakami et al.*, 2010], Sequential Self Calibration (SSC) [e.g., *Gómez-Hernández et al.*, 1997] or the quasi-linear geostatistical inversion approach [e.g., *Kitanidis*, 1995], iteratively correct individual model realizations until they meet the data.

All calibration methods, mentioned above, would need to solve an optimization problem for many realizations, just in order to account for a single possible set of measurement values from one single suggested sampling pattern. In addition, strong nonlinearities cause convergence problems of such linearized approaches. This leads to an infeasible computational load, because nonlinear data impact estimations requires repeating nonlinear inference for many (100...10,000) possible sets of measurement values per individual sampling pattern. This computational effort would again multiply with the number of different competing sampling patterns tested during the search for the optimal design (typically 1,000...1,000,000).

Bootstrap filter An alternative to parameter calibration is filtering or weighting given sets parameter. One flexible approach is the Bootstrap filter (BF) [*Gordon et al.*, 1993], a nonlinear and accurate conditioning method, that was introduced in Sec. 2.3.5. The BF does not invest any effort to correct individual realizations for meeting a given data set. Instead, the BF selects fitting realizations from the prior ensemble, where there goodness of fit is defined through their likelihood values. The likelihoods are expressed as weights, leading to a representation of the posterior by the weighted prior distribution. For conditioning on a different data set, only the weights will change, but no further iteration is necessary.

This weighting procedure would be an expensive procedure for the generation of one posterior state. However, the same prior can serve for different possible posterior states. So, the inference of many statistically stable posterior states requires only one sufficiently large prior as a source for fitting realizations. For this reason, repetition for (1) many possible sets of data values per

given design and (2) many possible designs to be compared would be relatively cheap. Therefore, this initially expensive filtering approach pays off combined with the need to average over many potential data sets. Due to these advantages, I will use the BF as the inference method. BFs are technically the same as the Bayesian Generalized Likelihood Uncertainty Estimator (Bayesian GLUE) by *Beven and Binley* [1992]; *Beven and Freer* [2001] when using the latter with formal (e.g., Gaussian) likelihoods. However, from a philosophical perspective, BF and GLUE pursue rather different paths, and the informal likelihoods commonly associated with GLUE are often being disputed in the scientific community [e.g., *Mantovan and Todini*, 2006].

4.2. Approach

This section introduces the extension of the Bootstrap Filter (BF) [*Gordon et al.*, 1993] towards estimation of data impact, i.e., for the optimal planning of data acquisition. Assuming that, for the current information state, the prior model is the best available estimator for the system, this framework uses one single ensemble for two MC approximation steps. The first step estimates the distribution of possible data sets for future measurements. The second step is the nonlinear inference by the BF, using the data likelihood leading to the conditional distribution $p(\mathbf{s}|\mathbf{y}_i(\mathbf{d}))$. Instead of $\mathbf{s}$, any other quantity of the model can be conditioned on a set of measurement values $\mathbf{y}_i(\mathbf{d}) \sim p(\mathbf{y}(\mathbf{d}))$. The distribution of measurements $p(\mathbf{y}(\mathbf{d}))$ is simulated by the current model and selected according to the proposed design d. The likelihood function may reflect both measurement and model errors. Under the assumption that both error types are multivariate Gaussian (typically uncorrelated), it is possible to analytically marginalize over possible measurement and model errors (see Sec. 4.3). This additional smoothing acts like a Kernel smoothing technique [e.g., *Wand and Jones*, 1995] and yields a substantial gain in computational time for the optimal design problem. The Gaussian assumption, moreover, poses no limitation to this methodological framework, since it can be replaced by any other preferred distribution. The analytical marginalization shown in Sec. 4.3.2 is a simple convolution between two error distributions that can be performed for many other combinations of parametric distributions of measurement and model errors.

Due to the underlying strictly numerical MC framework, the method does not suffer from linearity assumptions, and can handle all types of data in a similar fashion to *Zhang et al.* [2005] and *Murakami et al.* [2010]. Also, it is not limited to multi-normality assumptions of parameters like hydraulic conductivity, which are subject to increasing criticism [e.g., *Fogg et al.*, 1998; *Gomez-Hernandez and Wen*, 1998; *Bárdossy and Li*, 2008]. It opens the path to consider various sources of uncertainty, e.g., heterogeneity, geostatistical assumptions, boundary conditions, measurement values, model structure uncertainty or uncertain model choice, and model errors with known parametric distributions. This helps to minimize the strength or even necessity of possibly subjective prior assumptions on parameter distributions and modeling assumptions, which would be hard to defend prior to data collection [e.g., *Chaloner and Verdinelli*, 1995; *Diggle and Ribeiro*, 2007; *Nowak et al.*, 2009].

Depending on the task-specific context, the data utility $\phi(\mathbf{y}_i(\mathbf{d}))$ may directly measure the benefit of reduced parameter uncertainty in $\tilde{p}(\mathbf{s}|\mathbf{y}_i(\mathbf{d}))$, or include a prescribed prediction goal by

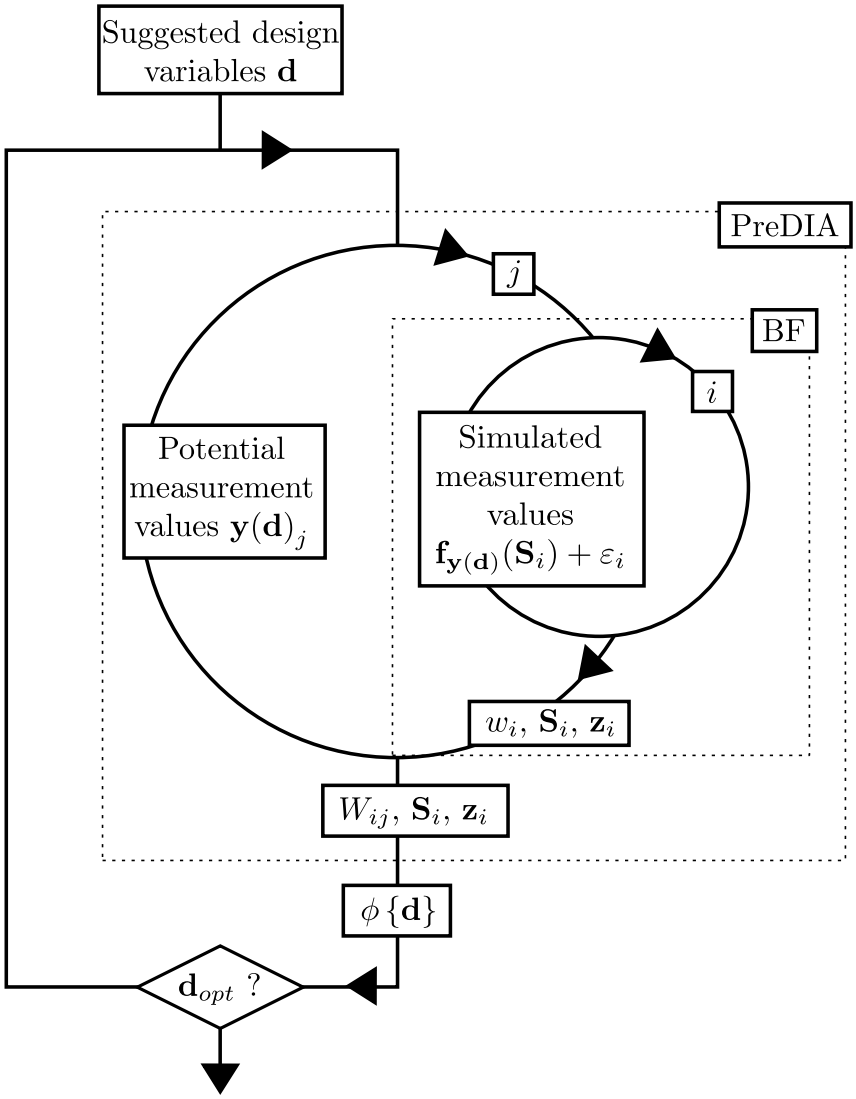


Figure 4.1.: Schematic illustration of PreDIA (Pre-posterior Data Impact Assessor) enveloping the BF (Bootstrap Filter). Both are nested together in the optimization procedure.

working on $\tilde{p}(\mathbf{z}|\mathbf{y}_i(\mathbf{d}))$ in a more complex manner. Without loss of generality and for reasons discussed in Sec. 3.3.2, I will focus on the latter option. Finally, I average the data utility function $\phi\{p(\mathbf{z}|\mathbf{y}_i(\mathbf{d}))\}$ (or analogously for the parameters) over the possible, yet unknown, measurement values via $\tilde{p}(\mathbf{y}(\mathbf{d}))$ to evaluate the design utility:

$$\Phi(\mathbf{d}) \propto \int\limits_{\mathbf{y}(\mathbf{d})} \phi(\tilde{p}(\mathbf{z}|\mathbf{y}(\mathbf{d})))\tilde{p}(\mathbf{y}(\mathbf{d}))\ \mathrm{d}\,\mathbf{y}(\mathbf{d}), \tag{4.1}$$

where

$$\tilde{p}(\mathbf{y}(\mathbf{d})) = \int\limits_{\mathbf{s},\boldsymbol{\xi},\mathbf{k},\boldsymbol{\theta}} L(\mathbf{y}(\mathbf{d})|\mathbf{s},\boldsymbol{\xi},\mathbf{k},\boldsymbol{\theta})\ p(\mathbf{s},\boldsymbol{\xi},\mathbf{k},\boldsymbol{\theta})\ \mathrm{d}\,(\mathbf{s},\boldsymbol{\xi},\mathbf{k},\boldsymbol{\theta})\,. \tag{4.2}$$

Each posterior state is a product of nonlinear Bayesian inference, with respect to a single anticipated set of data values. The number of considered posterior states needs to be sufficiently high to approximate the joint distribution $p(\mathbf{y})$ accurately, as the estimation quality for data impact estimates will heavily depend on it. The potentially large number of inference steps therefore requires very efficient concepts.

Eqs. (4.1)-(4.2) reveal that the prior distributions significantly influence the overall results. It is also well-known in BMA that its results are directly conditional on the set of models considered in the analysis. Therefore, the selection of model choices $\mathbf{k}$, structural parameters $\boldsymbol{\theta}$ and prior assumptions on all distributions should be as general as possible (least subjective), to reflect the situation prior to the investigation. Lack of information could, for example, result in a flat (least subjective) prior or even improper priors [e.g., *Kass and Wasserman*, 1996]. Please note that improper or flat priors in conjunction with BMA may cause severe problems [*Hoeting et al.*, 1999]. Maximum Entropy [*Jaynes*, 1957] or Minimum relative Entropy approaches [e.g., *Woodbury and Ulrych*, 1993; *Hou and Rubin*, 2005] can also be used to keep the prior description as general as possible. Another option is using reference priors [*Kass and Wasserman*, 1996]. The dependency on the prior and possibilities to reduce it will be the focus of Chap. 6.

Generally, the combination of the BF with a preposterior MC framework preserves any statistical features of arbitrarily high order, provided that the Monte-Carlo sample is large enough. However, it inherits the problem of high-dimensional filtering [e.g., *Liu et al.*, 2012].

4.3. Implementation

4.3.1. Nonlinear inference

The basis of preposterior data impact estimation is nonlinear Bayesian inference, which is executed using the bootstrap filter from Sec. 2.3.5. To simplify notation for this section, all individual parameter vectors $\mathbf{s}, \boldsymbol{\xi}, \boldsymbol{\theta}, \mathbf{k}$ are combined in the augmented vector $\mathbf{S}$. Gaussian error models are commonly used to simulate noisy data through $\mathbf{y}_i(\mathbf{d}) = \mathbf{f}_y(\mathbf{S}_i, \mathbf{d}) + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{\boldsymbol{\varepsilon}})$ with covariance matrix $\mathbf{R}_{\boldsymbol{\varepsilon}}$. In order to compute the likelihood of a realization i and a given data set $\mathbf{y}_0$, the BF would use the residual vector $\boldsymbol{\Delta}_0 = \mathbf{y}_0 - \mathbf{f}_y(\mathbf{S}_j)$ and, for n_y measurement values, set:

$$L(S_i|\mathbf{y}_0) = \frac{1}{\left((2\pi)^{n_y} \det \mathbf{R}_{\boldsymbol{\varepsilon}}\right)^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\Delta}_0\right)^T \mathbf{R}_{\boldsymbol{\varepsilon}}^{-1}\left(\boldsymbol{\Delta}_0\right)\right]. \tag{4.3}$$

Normalizing the likelihoods according to Bayes theorem yields the weight vector used in the BF:

$$w_i = \frac{L(\mathbf{S}_j|\mathbf{y_0})}{\sum_{j=1}^{n} L(\mathbf{S}_j|\mathbf{y_0})}, \tag{4.4}$$

which represents one posterior state based on the measurement data set $\mathbf{y}_0$.

4.3.2. Preposterior data impact assessment

The posterior conditioning on a hypothetically given vector of measurement values $\mathbf{y}_0$ is now extended towards the expected data impact of a given design $\mathbf{d}$, where the actual vector of measurement values is yet unknown. This situation reflects the utility evaluation of a single proposed design candidate $\mathbf{d}$ during the planning phase of DoE, called the preposterior stage [e.g., *James and Gorelick*, 1994]. The prior model is used to extract m measurement vectors $\mathbf{y_j}(\mathbf{d})$, which are drawn according to their prior distribution $\tilde{p}(\mathbf{y}(\mathbf{d}))$ as defined in Eq. (4.2). Expected data impact is also known under the term Expected Value of Measurement Information (EVMI), which was previously introduced in Sec. 3.3.

Likelihood estimation: The evaluation of the preposterior state is computationally demanding when no specific measures are taken (see Sec. 4.3.3). The generation of multiple *potential* data values $\mathbf{y}_j$,which simulate the distribution $p(\mathbf{y}(\mathbf{d}))$, uses as well the simulation model $\mathbf{y}_j = \mathbf{f}_y(\mathbf{S}_j) + \boldsymbol{\varepsilon}_j$ with an additional Gaussian error term $\boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{\boldsymbol{\varepsilon}})$. This yields

$$L(\mathbf{S}_i|\mathbf{S}_j, \boldsymbol{\varepsilon}_j) = \frac{1}{((2\pi)^{n_y} \det \mathbf{R}_{\boldsymbol{\varepsilon}})^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\Delta}_j + \boldsymbol{\varepsilon}_j\right)^T \mathbf{R}_{\boldsymbol{\varepsilon}}^{-1}\left(\boldsymbol{\Delta}_j + \boldsymbol{\varepsilon}_j\right)\right], \tag{4.5}$$

with $\boldsymbol{\Delta}_j = \mathbf{f}_y(\mathbf{S}_j) - \mathbf{f}_y(\mathbf{S}_i)$. It is further possible to average over all *potential* measurement error values of $\boldsymbol{\varepsilon}_j$ via the marginalization $L(\mathbf{S}_i|\mathbf{S}_j) = \int_{-\infty}^{+\infty} p(\mathbf{S}_i|\mathbf{S}_j, \boldsymbol{\varepsilon}_j) \cdot p(\boldsymbol{\varepsilon}_j)\ \mathrm{d}\,\boldsymbol{\varepsilon}_j$.

Derivation of the weight matrices: For a vector of m possible sets of measurement values, each one leading to a $n \times 1$ weight vector $\mathbf{w}_j$, this yields an $n \times m$ weight matrix $\mathbf{W}$. A schematic illustration of this procedure is shown in Fig. 4.1. Normalizing according to Bayes theorem yields:

$$w_{ij} = \frac{L(\mathbf{S}_i|\mathbf{S}_j)}{\sum_{i=1}^{n} L(\mathbf{S}_i|\mathbf{S}_j)}. \tag{4.6}$$

Preposterior statistics: Now, weighted averaging over n realizations and m possible vectors of measurement values yields the expected conditional prediction variance. The variance is computed using the fast computing formula after [e.g., *Weiss*, 2006, p. 355]:

$$E_{\mathbf{y}(\mathbf{d})}\left\{V_{\mathbf{z}|\mathbf{y}(\mathbf{d})}\left[\mathbf{z}\right]\right\} \approx \frac{1}{m}\sum_{j=1}^{m} \frac{v_{1,j}}{v_{1,j}^2 - v_{2,j}} \left\{\sum_{i=1}^{n} \mathbf{z}_i^2 \cdot w_{ij} - \left(\sum_{i=1}^{n} \mathbf{z}_i \cdot w_{ij}\right)^2\right\}, \tag{4.7}$$

which is significantly faster than other computation formula. This allows to compute the nonlinear utility function based on the variance as:

$$\Phi_{forw\ var}(\mathbf{d}) = E_{\mathbf{y}(\mathbf{d})}\left\{V[z] - V[z|\mathbf{y}(\mathbf{d})]\right\}. \tag{4.8}$$

In a similar fashion, other utility functions that act on $p(\mathbf{z}|\mathbf{y}(\mathbf{d}))$ can be defined as introduced in later parts of this thesis. The alert reader may have noticed that Eq. (4.7) might consume large CPU resources. This is caused by the need to average over the yet unknown measurement values. Hence, I suggest to keep the sample size for potential measurement values m much smaller than for potential measurement simulations n, since an accurate expectation over $p(\mathbf{y}(\mathbf{d}))$ requires less realizations than the conditional variance of $\mathbf{z}$ (also see Eq. (4.4)).

4.3.3. Efficient implementation within PreDIA

The PreDIA framework employs two concepts to increase in efficiency of the data impact assessment. Firstly, it is possible to use the statistical independence between $\boldsymbol{\varepsilon}_j$ and $\boldsymbol{\varepsilon}_i$ to analytically average over the two measurement errors, which results in:

$$L(\mathbf{S}_i|\mathbf{S}_j) = \frac{1}{\left((2\pi)^{n_y} \det \mathbf{R}_{\boldsymbol{\varepsilon}}^*\right)^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\Delta}_j\right)^T \mathbf{R}_{\boldsymbol{\varepsilon}}^{*-1}\left(\boldsymbol{\Delta}_j\right)\right], \tag{4.9}$$

with $\boldsymbol{\Delta}_j = \mathbf{f}_y(\mathbf{S}_j) - \mathbf{f}_y(\mathbf{S}_i)$ and $\mathbf{R}_{\boldsymbol{\varepsilon}}^* = 2\mathbf{R}_{\boldsymbol{\varepsilon}}$. This is equivalent to using noise-free potential data, but doubling $\mathbf{R}_{\boldsymbol{\varepsilon}}$ in the likelihood analysis. This saves the numerical simulation of ϵ_j and leads to a broader likelihood function, which partly attenuates filter degeneration.

The extension to model structure error can follow the same procedure like incorporating measurement errors: Both synthetic and potential data are then simulated by $\boldsymbol{y}_j = \mathbf{f}_y(\mathbf{S}_j) + \boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}_j^m$ and $\boldsymbol{y}_k = \mathbf{f}_y(\mathbf{S}_k) + \boldsymbol{\varepsilon}_k + \boldsymbol{\varepsilon}_k^m$. Following the precisely identical procedure, and if $\boldsymbol{\varepsilon}^m$ is assumed to be Gaussian with $\boldsymbol{\varepsilon}^m \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{\boldsymbol{\varepsilon}}^m)$ and model error covariance matrix $\mathbf{R}_{\boldsymbol{\varepsilon}}^m$, it can again be absorbed into the error covariance matrix $\mathbf{R}_{\boldsymbol{\varepsilon}}$. Therefore, it becomes $\mathbf{R}_{\boldsymbol{\varepsilon}}^* = 2\mathbf{R}_{\boldsymbol{\varepsilon}} + 2\mathbf{R}\boldsymbol{\varepsilon}^m$.

Secondly, as the model is the best available predictor for $p(\mathbf{y})$, represented by an ensemble, this ensemble is at the same time used to extract potential data sets $f_y(\mathbf{S}_j)$ and the parameter vector $\mathbf{S}_i$. In formal Bayesian updating, the likelihood L_{ij} of parameter vector $\mathbf{S}_i$ given the synthetic noise-free $n_y \times 1$ vector of measurement values $\mathbf{f}_y(\mathbf{S}_j)$ resembles a multivariate normal distribution in $\Delta = \mathbf{f}_y(\mathbf{S}_j) - \mathbf{f}_y(\mathbf{S}_i)$ with mean zero and covariance matrix $\mathbf{R}_{\boldsymbol{\varepsilon}}{}^* = 2\mathbf{R}_{\boldsymbol{\varepsilon}}$, where $\mathbf{R}_{\boldsymbol{\varepsilon}}$ is the covariance matrix of measurement errors. Hence:

$$L_{ij} = \frac{1}{\left((2\pi)^{n_y} \det \mathbf{R}_{\boldsymbol{\varepsilon}}^*\right)^{1/2}} \exp\left[-\frac{1}{2}\left(\Delta_j\right)^T \mathbf{R}_{\boldsymbol{\varepsilon}}^{*-1}\left(\Delta_j\right)\right] \;\forall\; i \neq j, \tag{4.10}$$

$$L_{ij} = 0 \;\forall\; i = j. \tag{4.11}$$

Since, for $i = j$, $\mathbf{S}_i$ and $\mathbf{S}_j$ would not be drawn independently, the evaluation is discarded $i = j$ and instead the likelihoods L_{ii} are set to zero. This procedure remains the choice of the modeler and has advantages and drawbacks.

4.4. Convergence

The convergence of PreDIA is composed of two (partly independent) converging dimensions: The first considers the stochastic accuracy of the individual filtered posterior states. It is related to the accurate approximation of the posterior statistics (e.g., variance) for a given hypothetical data set $\mathbf{y}_0$. The second is related to the accurate stochastic representation of the distribution of possible measurements $p(\mathbf{y})$ and is therefore connected with the number of considered posterior states and the resulting *expected* value $E_y(\cdot)$ of the applied uncertainty measure. I will discuss both parts in this section.

4.4.1. Filter convergence

In order to obtain robust Monte Carlo (MC) statistics, a sufficient amount of realizations is needed. The convergence of the individual posterior states depends on the number of realizations that represent the posterior, e.g., the number of weighted realizations that effectively remain after filtering. Obviously, PreDIA inherits the curse of dimensionality from other filtering techniques. However, the problem of filter degeneracy is mitigated in PreDIA to a substantial extent because of the analytical marginalization of the likelihood function over measurement and model errors of predicted potential data values. This widens the likelihood, acts like a kernel smoothing technique [e.g., *Wand and Jones*, 1995], and so averages out statistical noise in conditional variances. It is important to understand that this marginalization is not an assumption or approximation that would artificially weaken the data or compromise the Bayesian framework. It is merely an analytical step within an otherwise numerical framework that takes advantage of the two Gaussian distributions. The kernel smoothing properties of this step are a crucial aspect of PreDIA, because informative and low-error data sets are exactly what one desires to have, and to optimize in applications.

In order to quantify the remaining degree of filter degeneracy of PreDIA, the Effective Sample Size (see Sec. 2.3.5) is evaluated and averaged across all possible data values to obtain the averaged effective sample size (AESS):

$$AESS = \frac{1}{m}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}(w_{ij})^2\right)^{-1}. \tag{4.12}$$

The AESS is recorded it during the optimization procedure. Values that are too low indicate that the current analysis requires a larger prior ensemble. Re-sampling strategies do not make sense in the context of PreDIA, because resampling from all possible posterior distributions is equivalent to enlarging the entire sample right from the start.

The AESS cannot effectively ensure that all individual ESS values exceed a required level. Alternatively, a minimum required ESS value for all posterior states could be defined. However, by doing so, data realization leading to the word individual filter degeneration would define the required ensemble size for all realizations. For all other states, the minimum would be exceeded. Thus, only an adaptive approach to select the ensemble size individually in each state would be equally efficient in all states. While this would increase the processing speed, yet the minimum required ESS value would define the overall required number of simulated realizations.

4.4.2. Preposterior convergence

It remains to investigate how well the preposterior states and ther relevant averaged utility value converges. Eq. (4.1) shows that averaging over potential preposterior states is basically an integration over the joint distribution of measurements $p(\mathbf{y})$. This is executed by a Monte-Carlo integration and therefore the same convergence rate of MC integration as introduced in Sec. 2.2.4 applies. The typical convergence rate is $O(1/\sqrt{(}N))$, where N is the sample size

[e.g., *Caflisch*, 1998; *Ballio and Guadagnini*, 2004]. Therefore, the required number of considered posterior states heavily depends on the variance and shape of the distribution of data impact $\phi(\mathbf{y}(\mathbf{d}))$ over the distribution $\phi(\mathbf{y}(\mathbf{d}))$ and the desired accuracy. *Diggle and Lophaven* [2006] and *Neuman et al.* [2012], for example, found empirically that their statistics stabilize at $n = 100$ or $n = 200$ sets of randomly drawn data sets for averaging conditional variances. Variance reduction methods [e.g., *Russell*, 1986; *Newman and Barkema*, 1999; *Guthke and Bárdossy*, 2012; *Cheng*, 1986] such as stratified sampling, importance sampling or antithetic sampling can help to further improve the situation.

During this study, it was observed that the averaging over many preposterior states also has a smoothing effect on the noise introduced in the filter due to small (A)ESS values. However, a bias effects introduced by the filtering collapse are not mitigated. Thus, both parts of convergence need to be ensured separately, but are not entirely independent.

4.4.3. Influence factors for convergence

The more freedom in model and parameter choice is admitted, the more complex the model may become. As a general property of MC, this does not affect the convergence rate of MC statistics, but the involved statistical distributions (e.g., heavy tails) and dimensionality affect the initial uncertainty of the estimate. Therefore, the convergence properties of the BF and related methods may depend on the variance and higher moments of simulated data, but do not generally depend on model complexity in a direct fashion. However, when looking at more complex problems, two additional problems may arise through indirect mechanisms.

First, multivariate structure of possible data may become more complex under more complex models. For example, when sampling only hydraulic conductivity or hydraulic heads in 3D versus 2D, their spatial variations remain, in principle, the same. In such cases, the number of required realizations for MC, BF or PreDIA does not increase. For concentration measurements, however, the multivariate structure in 3D is much more complex than in 2D, because a transported solute plume has more spatial degrees of freedom in 3D. In such cases, the required number of realizations may increase at a higher rate, as the number of considered measurements increase.

Second, the design space increases when switching from 2D to 3D or when optimizing the schedule of an experimental design for dynamic systems rather than static patterns. With more potential measurement locations, the burden of the high-dimensional optimization problem in Eq. (3.1) increases. This is, however, not an artifact of the proposed PreDIA method, but rather a general problem shared by all OD methods. This problem will require more future research on adequate optimization algorithms.

As for implicit BMA, it is important to stress that PreDIA can principally handle uncertain discrete choice between structurally different model alternatives at no additional conceptional costs within the data analysis (again see Sec. 2.3.4). The method merely requires convergence of the overall (combined) sample statistics and not necessarily individual convergence for each single considered model.

4.5. Application for a groundwater scenario

To demonstrate the properties, advantages and relevance of nonlinear data impact estimation as done by PreDIA, a synthetic application scenario was set up. The scenario is taken from the area of groundwater contamination problems. I choose an example from subsurface contaminant transport, since this allows to illustrate aspects of model choice and uncertain boundary conditions, and because subsurface transport reveals strongly nonlinear and yet well-researched dependencies between predictions and parameters.

4.5.1. Scenario definition and configuration

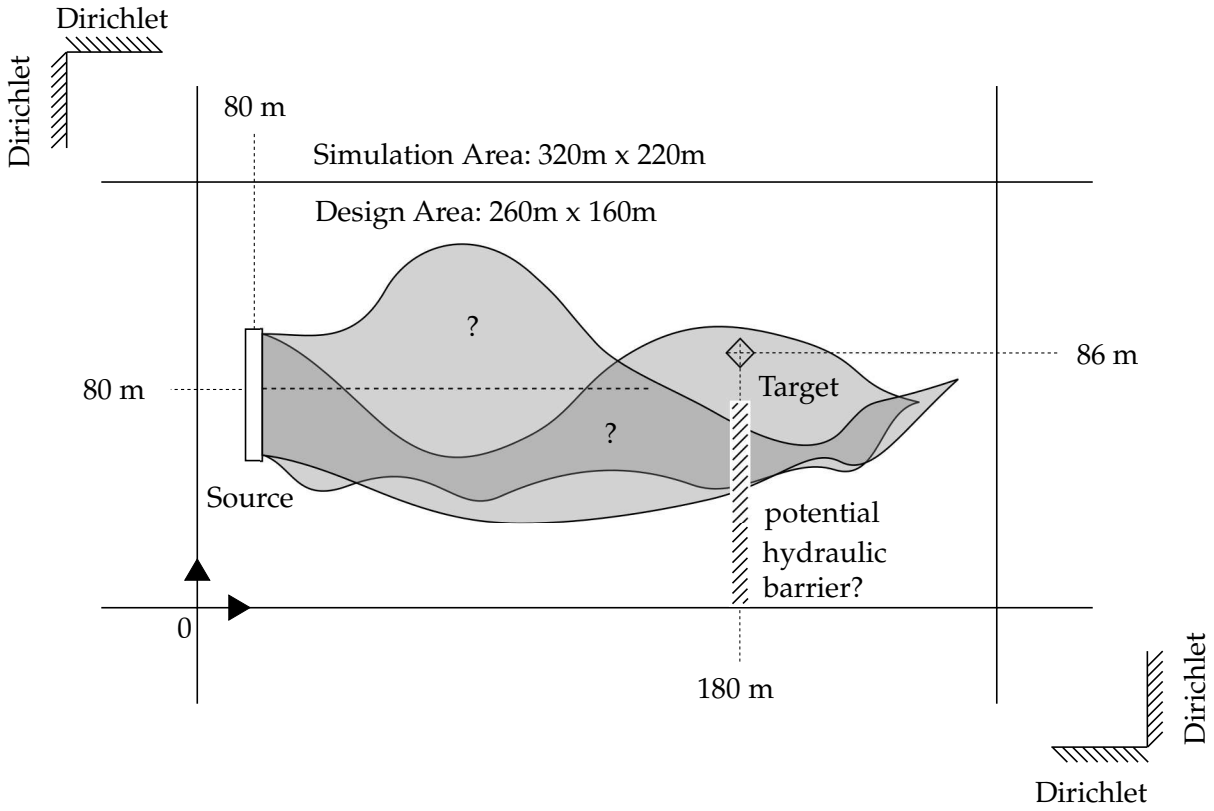


Figure 4.2.: Spatial configuration of the synthetic test case involving a contaminant release at (x_S, y_S) = (80m, 80m) with source width l_s = 20m, prediction target at (x_{tr}, y_{tr}) = (180m, 86m), and a possibly present hydraulic barrier due to a different geological zonation (considered in application case 3). For further details, see Tab. 4.1.

The synthetic scenario assumes a drinking water well or a similar sensitive location threatened by a recent but continuous contaminant source located upstream. The goal is to find the sampling pattern which optimally reduces the uncertainty of predicting the long-term (steady

state) contaminant concentration to be expected at the sensitive location. For simplicity, but not as a limitation of the methodology, the test case considers steady-state groundwater flow in a 2D depth-averaged heterogeneous aquifer as described in Sec. 2.1.1. Transport boundary conditions are specified as Dirichlet boundaries with $\hat{c}(x) = 0$ at all outer boundaries and $\hat{c}(x) = c_0$ at the source. For simplicity, the highest concentrations (at the source) are normalized to $c_0 = 1$. All known flow- and transport parameters and their values are summarized in Tab. 4.1. An overview of the domain geometry including the source and the sensitive location is provided in Fig. 4.2. The source location (x_S, y_S) = (80m, 80m) and its width l_s = 20m are assumed to be known. The sensitive location is at (x_{tr}, y_{tr}) = (180m, 86m). Within the geostatistical setup provided below, this is about seven expected-integral scales downstream of the contaminant source and about half an integral scale offset from the center line of the expected plume path.

Fig. 4.2 also depicts a possibly present hydraulic barrier (width = 10 m) due to uncertainty in geological medium boundaries. To illustrate structural model uncertainty, I assume that local hydrogeologists are uncertain about the extent of a narrow zone filled with a different geological facies which might be present in that area. For simplicity, this is implement as a rectangle with a different mean value of log-conductivity at $T(\mathbf{x})$ = ln 10^{-7}. The prior probability of this alternative model is set to 30%. Please note that the possibly present barrier is only considered in the last test case (Sec. 4.5.3).

In the example, the global transmissivity field is uncertain, following a geostatistical approach. As an example for structural model uncertainty, the geostatistical model will be kept uncertain. Bayesian geostatistics, as introduced in Sec. 2.3.3, open the path to a more general consideration of uncertainty. Let $\boldsymbol{\theta} = (\sigma^2, \lambda_i, \kappa)$ contain uncertain structural parameters of the geostatistical model, where σ^2 accounts for the field variance, λ_i are the correlation length scales in spatial directions x_i and κ is the shape parameter of the Matérn covariance function (see Sec. 2.3.4).

Furthermore, I assume the flow boundary conditions to be uncertain as well. The values for the Dirichlet conditions are determined by the uncertain angle ν which defines the regional head gradient orientation relative to the northern/southern boundaries. All uncertain geostatistical parameters $\boldsymbol{\theta}$, boundary parameters $\boldsymbol{\xi}$ and conceptual models $\boldsymbol{k}$ are drawn from their respective distributions (see Tab. 4.2).

Concentrations c are considered to be not available as measurement data, because the spill just happened and the plume has not evolved yet. Instead, hydraulic head and transmissivity data shall be optimally collected in order to minimize the prediction variance associated with the prediction. Hence, Eq. (4.8) is the utility function for this task. Measurements of transmissivity T and hydraulic head h are subject to measurement errors $\sigma_{r,T}$ and $\sigma_{r,h}$, respectively, since they are measurable only at the point scale and in an imprecise manner, e.g., by disturbed core-samples and small monitoring wells. For instructive reasons, I decide not to sample transmissivity T at the same locations as hydraulic head h by default, since this will help to better exhibit and discuss the underlying physics associated with the respective choice of location and data type. Locations where T is informative may not be informative for h measurements, because different physical flow and transport-related phenomena may co-ordinate the individual data types to different informative locations. However, the optimization framework could easily handle constraints such that T and h measurement locations have to coincide.

Numerical domain			
Domain size	$[L_1, L_2]$	$[m]$	[320, 220]
Grid spacing	$[\Delta_1, \Delta_2]$	$[m]$	[0.25, 0.125]
Design domain			
Domain size	$[L_1, L_2]$	$[m]$	[260, 160]
Grid spacing	$[\Delta_1, \Delta_2]$	$[m]$	[2, 2]
Transport parameters			
Head gradient	γ	[-]	0.01
Effective porosity	n_e	[-]	0.35
Local-scale dispersivities	$[\alpha_l, \alpha_t]$	$[m]$	[0.5, 0.125]
Diffusion coefficient	D_m	$[m^2/s]$	10^{-9}
Transversal plume dimension	l_s	$[m]$	20
Known geostatistical model parameters			
Global mean	$\beta_1 = \ln T$	[-]	$\ln 10^{-5}$
Trend in mean	β_2	[-]	0
Measurement error standard deviations			
Hydraulic conductivity	$\sigma_{r,T}$	[-]	1.00
Hydraulic head	$\sigma_{r,h}$	[m]	0.01

Table 4.1.: Known parameters for the flow, transport and geostatistical model.

Uncertain structural parameters $\boldsymbol{\theta}$			
Variance	σ_T^2	[-]	$\mathcal{N}(\mu = 2.0, \sigma = 0.3)$
Integral scale	λ	[m]	$\mathcal{N}(\mu = 15, \sigma = 2.0)$
Matérn Kappa	κ	[-]	$\mathcal{U}(a = 5, b = 36)$
Uncertain boundary parameters $\boldsymbol{\xi}$			
Deviation from center	ν	[°]	$\mathcal{N}(\mu = 0.0, \sigma = 10)$
Uncertain conceptual models $\boldsymbol{k}$			
Existence of hydraulic barrier	-	[-]	$\mathcal{B}(p = 0.3)$

Table 4.2.: Uncertain structural and boundary parameters and their assigned distributions.

The implementation is done in the MATLAB environment. For generating geostatistical random fields and simulating groundwater flow and solute transport, I use the same code already used in *Nowak et al.* [2008] and *Nowak et al.* [2009]. A sample size of $n = 70,000$ realizations has been chosen to ensure that the discussion of the method and resulting designs is not compromised by statistical noise. The sequential exchange algorithm (see Sec. 3.2.1) is used in order to optimize the design, and the utility of each design candidate is evaluated with PreDIA, based on Eq. (4.8) as the utility function.

4.5.2. Scenario variations

Two different scenario objectives are considered. Both serve to illustrate how arbitrary prediction goals (regardless of their nonlinearity) can be handled within PreDIA and together with the ability to include arbitrary task-driven formulations. A third scenario is then exclusively addressing the consideration of conceptual model uncertainty, i.e., via incorporating a hydraulic barrier:

- Scenario 1: Minimum-variance prediction of contaminant concentration c at the sensitive location. To emphasize the difference to conventional linear methods, I compare the results of PreDIA to results from an Ensemble Kalman Filter (EnKF) [e.g., *Herrera and Pinder*, 2005; *Evensen*, 2007]. Therefore, the first scenario (case 1a) employs PreDIA for data impact estimation, whereas a scenario variation (case 1b) uses the EnKF (case 1b).
- Scenario 2: Maximum-confidence prediction of whether a critical concentration threshold will be exceeded at the sensitive location. This is equivalent to predicting an indicator quantity $z = I(c > c_{crit})$, with $E[I] = P(c > c_{crit})$. Since the indicator is a discrete variable that depends very nonlinearly on model parameters, it does not meet the requirements under which EnKFs can be used for comparison. Instead, two threshold values are considered with PreDIA: $c_{crit} = P_{15}$ (case 2a) and $c_{crit} = P_{75}$ (case 2b), where P_{15} and P_{75} are the c-values below which 15 % and 75 % of the c-values may be found according to the used prior distribution, respectively.
- Scenario 1: Consideration of a hydraulic barrier and minimum-variance prediction of a contaminant concentration c at the sensitive location (case 3).

The following sections present and discuss the sampling patterns resulting from the synthetic test cases and their variations defined in the previous section.

Fig. 4.3 shows the variances of transmissivity T (top), hydraulic head h (middle) and concentration c (bottom) prior to investigation and thence at the initial point for the considered DoE effort.

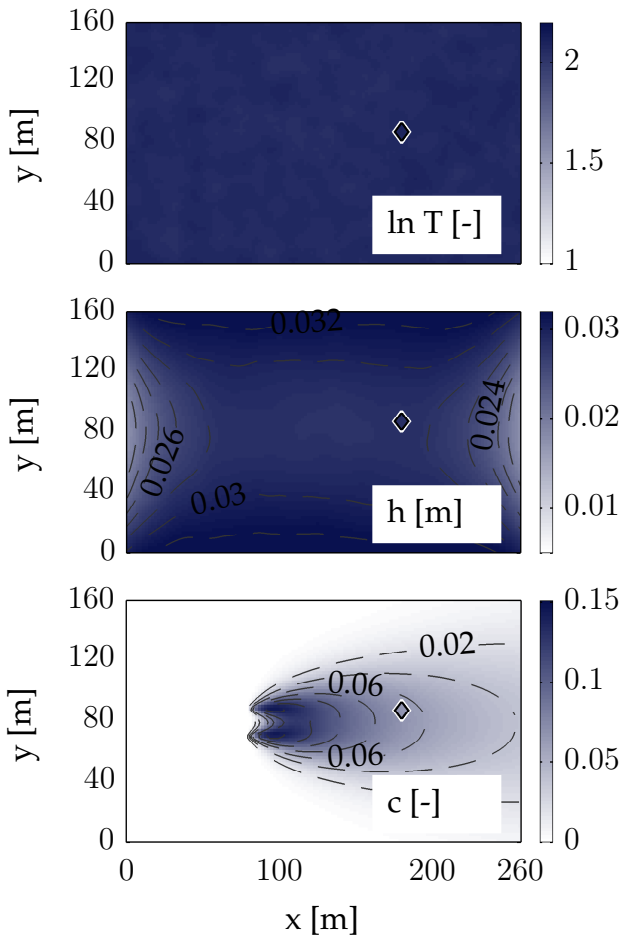


Figure 4.3.: Prior uncertainties (variance) associated with transmissivity (top), hydraulic head (center) and concentration (bottom) based on the uncertain structural and boundary parameters listed in Tabs. 4.1-4.2.

4.5.3. Results and discussion

Case 1a: Sampling patterns

Case 1a features optimal sampling for minimum-variance prediction of concentration at the sensitive location. The resulting sampling pattern, obtained with PreDIA, is shown in Fig. 4.4. Fig. 4.4 also includes the expected conditional variance of transmissivity (top), hydraulic head (center) and predicted concentration (bottom) according to Eq. (4.7). The basic characteristics of the design pattern mostly coincide with the results found in *Nowak et al.* [2009] who considered a similar scenario. However, there are important differences since they used the linear EnKF and I employ nonlinear estimation methods. With regard to the sampling pattern, I separate two predominant groups: (1) measurements gathering around the source and (2) measurements flanking the expected migration path of the plume.

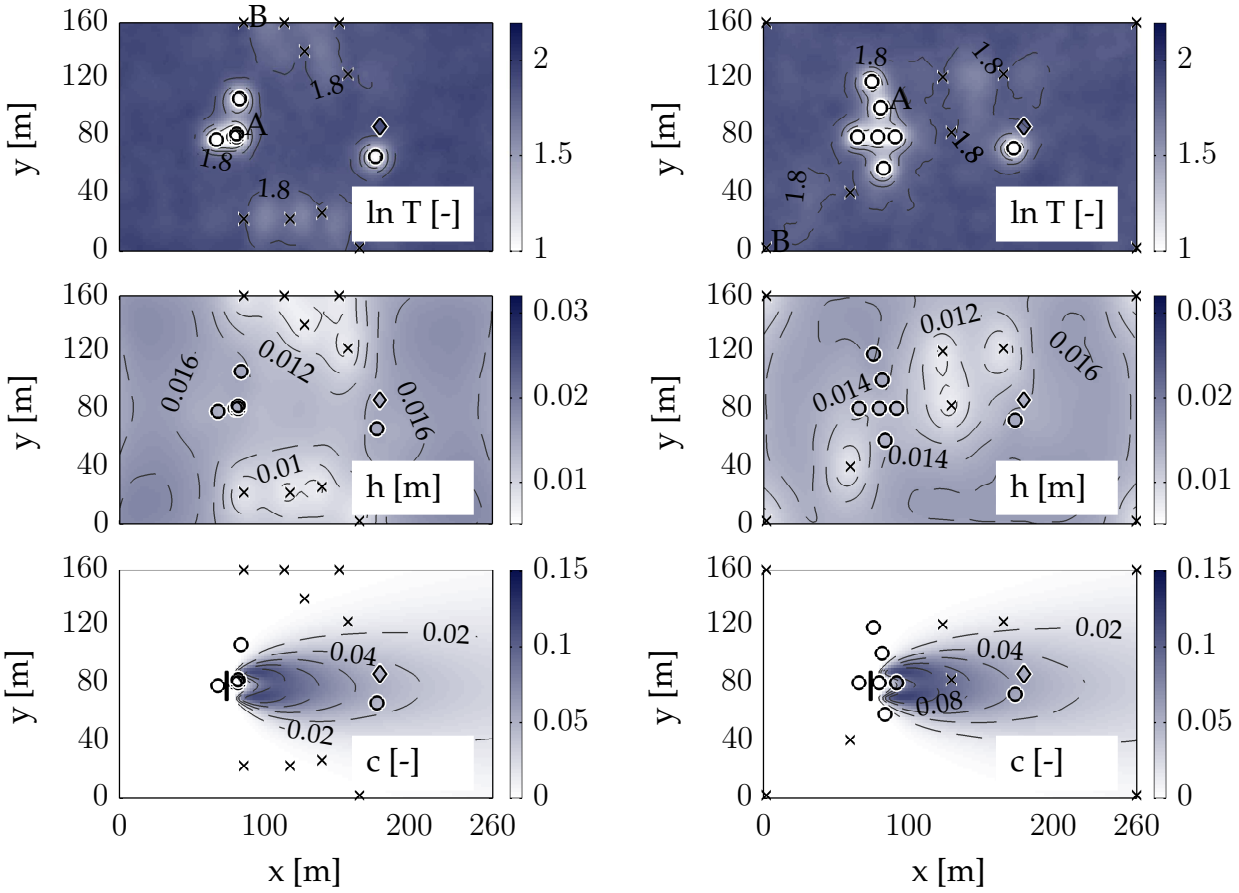


Figure 4.4.: PreDIA-based (left, case 1a) and EnKF-based (right, case 1b) sampling pattern optimized for minimum prediction variance of concentration at the sensitive location. Head measurements (crosses), transmissivity measurements (circles), source (box) and target (diamond). Maps in the background are expected preposterior variances for transmissivity (top), hydraulic head (center) and concentration (bottom).

Near-source measurements are exclusively occurring as transmissivity measurements. They are highly informative since they provide information about the volumetric flow rate through the source area. The flow rate through the source, in turn, is a dominant factor that dictates the total contaminant mass flux, the expected width, and the dispersion characteristics of the plume further downstream [*de Barros and Nowak*, 2010].

The measurements flanking the plume are head measurements which capture both the large-scale drift of the plume (due to the uncertain regional head gradients) and the meso-scale meandering of the plume (caused by heterogeneity).

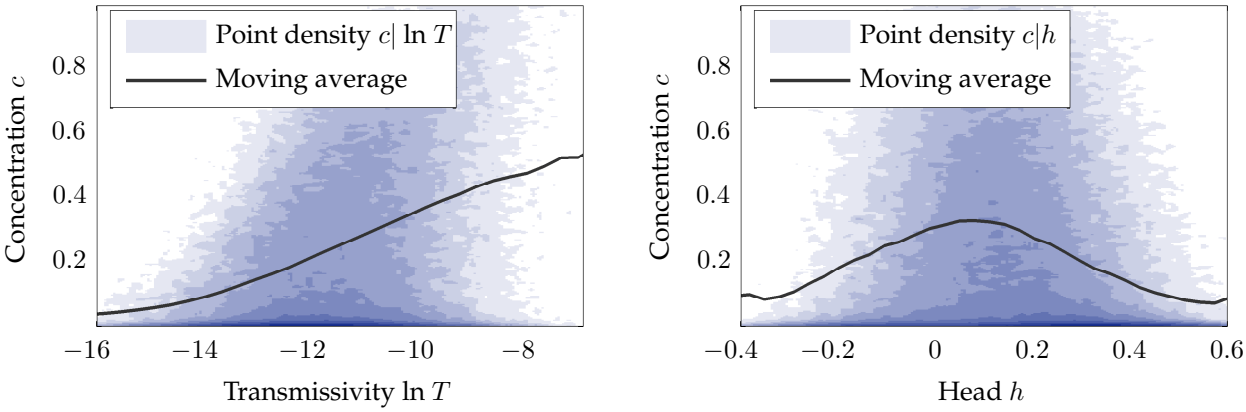


Figure 4.5.: Scatter density plots depicting the relation between the sample of predicted concentrations and the sample of transmissivity values at a near-source location A (left) and hydraulic head values at a near-boundary location B (right). The solid line illustrates the relation via moving average.

Statistical dependencies

In principle, the prediction task leads to information needs mostly in those regions where the statistical dependency between the measurable quantities (transmissivity or hydraulic head) and the prediction goal is highest, while avoiding mutually too close measurements that would merely convey to redundant information. Fig. 4.5 shows the statistical dependencies between observable quantities at potential measurement locations and the prediction target for a near-source transmissivity measurement location (A) in Fig. 4.4 and for a near-boundary head measurement location (B) in Fig. 4.4. The statistical dependencies are obtained by plotting the sample of possible measurement values against the sample of predicted concentrations. I additionally illustrate the nonlinear dependency in the scatter plot by a moving average line.

Location A: Obviously, transmissivity T at the near-source location (A) has a mostly linear relation to the predicted concentration. The higher values of the transmissivity at the source lead, on average, to higher source discharge and hence to a broader downstream plume after leaving the source. Therefore, the plume is far more likely to maintain high concentrations even over long travel distances, and is more likely to hit the target [*de Barros and Nowak*, 2010].

Location B: Opposed to that, head measurements h at the near-boundary location (B) exhibit a nonlinear dependency to the prediction goal. Extreme angles of the regional flow gradient divert the plume away from the target location, for both positive and negative values of the angle. By contrast, regional flow in the straight uniform direction, most likely, drives the plume through the target. The resulting dependency between hydraulic heads close to the boundary and the predicted concentration has an almost purely quadratic behavior, and shows almost no correlation in a linear sense, i.e. has almost zero covariance.

Reduction of uncertainty

Fig. 4.6 (left) illustrates how the individual transmissivity or hydraulic head measurements added during sequential design reduce the variance of the prediction goal and related physical quantities. The latter include the total solute mass flux through the source, the angle of the boundary condition (causing a large-scale drift), the width of the plume at the target (lateral spreading) and the lateral position of the plume's centroid (also affected by meso-scale meandering caused by heterogeneity).

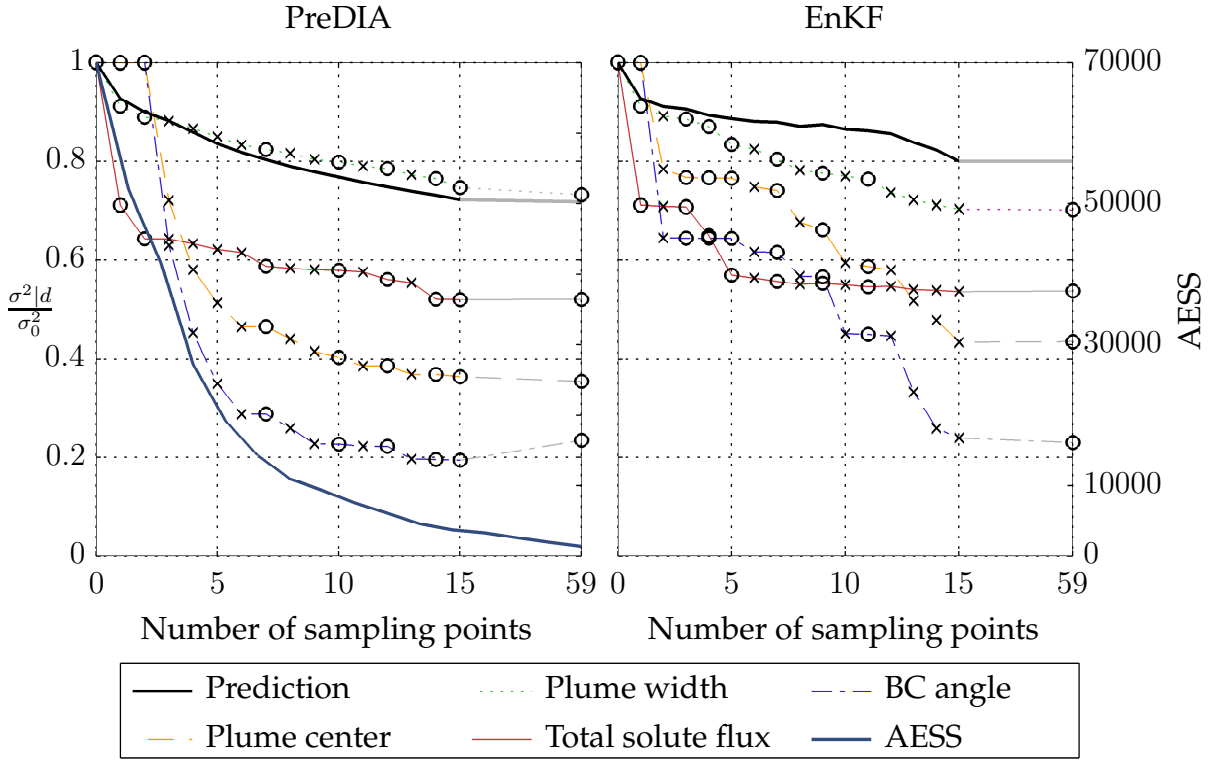


Figure 4.6.: Expected variance reduction (Eq. (4.7)) for PreDIA (left, case 1a) and EnKF (right, case 1b) during initial placement of samples for different auxiliary quantities. The sequential exchange phase is not shown in detail, but only indicated by the gray lines. Hydraulic head measurements are denoted by cross marks and transmissivity measurements by circle marks. The right axis quantifies, for the PreDIA-based optimization, the respective Averaged Effective Sample Size (AESS).

One clearly sees that transmissivity measurements located closely to the source greatly reduce the prediction uncertainty of the total solute flux (also see Fig. 4.6) for this case, while the head measurements along the flanks are almost uninformative to the total solute flux. Instead, the uncertainty of the boundary condition (regional flow direction) is greatly reduced by the head measurements, whereas the transmissivity measurements around the source contribute almost no related information (also see Fig. 4.6). Likewise, the position of the plume center is revealed almost solely by head measurements. The plume width at the prediction target is sensitive to

both head and transmissivity measurements, where the first two transmissivity measurements at the source are clearly the most valuable ones.

As discussed in Sec. 4.4, the Averaged Effective Sample Size (AESS) is one possible measure to monitor and avoid filter degeneracy and is surveyed during the optimization procedure. Fig. 4.6 (left) indicates the AESS (scale shown on the right axis) during the optimization scheme. The AESS drops from initially 70,000 to about 500. This is fully sufficient in order to calculate noise-free maps of the expected conditional variance (see Fig. 4.6) and evaluate the objective function reliably.

Comparison to EnKF (Case 1b): Sampling patterns

The sampling pattern provided by the Ensemble Kalman Filter (EnKF) relies on exactly the same geostatistical and boundary parameters used in case 1a, and hence uses the very same sample data. For technical insights into the EnKF formalism, please refer to *Herrera and Pinder* [2005] or *Evensen* [2007]. The resulting pattern is shown in Fig. 4.4 (right column). The underlaid maps of expected conditional variance are evaluated by PreDIA, because the maps provided by the EnKF are inaccurate and would not be comparable to those shown in the left part of Fig. 4.4.

Compared to the PreDIA-based sampling pattern (case 1a), one can find again the group of transmissivity samples in the source area. However, the number of measurements in this group is much larger. The next fundamental difference to the PreDIA-based sampling pattern is that the group of head measurements at the northern and southern domain boundary is smaller in favor of head measurements in the corners of the design domain. Apparently, the relevance of the variable boundary conditions that induce large-scale ambient flow deviation is also recognized, but judged differently by the EnKF analysis scheme.

The EnKF assesses statistical dependencies only via covariances, which are a measure for linear dependence only. It is unable to capture even-order (e.g. quadratic) dependencies such as between head measurements near the northern and southern boundary and the prediction goal (see Fig. 4.5). Therefore, it simply ignores these head measurement locations as potential sources of valuable information. Hence, crucial information about the meso-scale meandering of the plume is neglected. However, four measurement locations were placed at the corners of the allowable design locations. Apparently, their nonlinear dependency exhibits a sufficiently large linear component due to the slight asymmetry of the setup.

Direct comparison of data impact

For the sequential placement of measurements, data impact maps for each added measurement can be compared individually. A data impact map plots the value of data impact for all possible data points at their respective location. Fig. 4.8 compares the data impact maps for all steps of

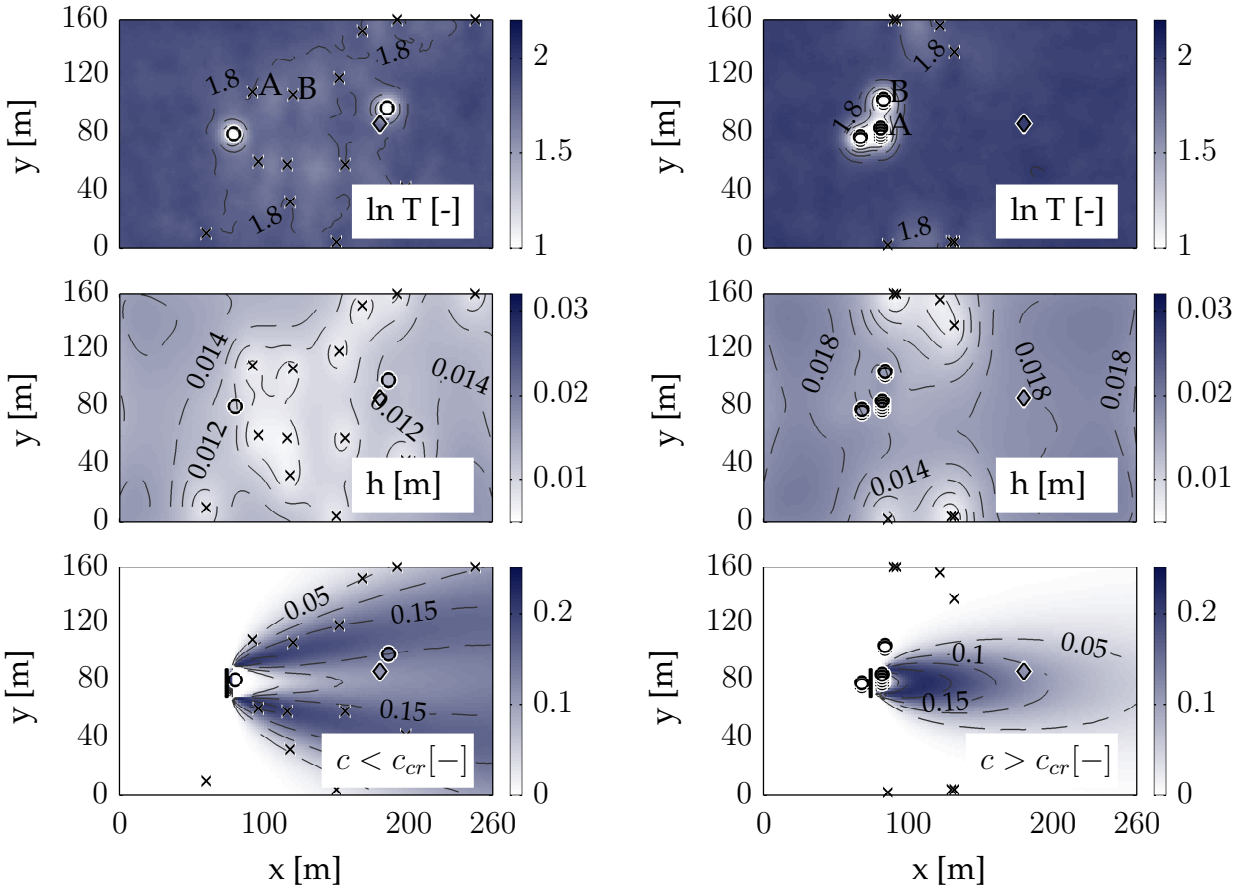


Figure 4.7.: PreDIA-based sampling pattern optimized for predicting the exceedance of a low c_{crit} (left, case 2a) and high c_{crit} (right, case 2b). Head measurements (crosses), transmissivity measurements (circles), source (box) and target (diamond). Maps in the background are preposterior variances for transmissivity (top), hydraulic head (center) and indicator variable (bottom). A selected near-source location is marked by A, whereas a near-boundary location is marked by B.

the sequential designs. The scatter plot assigns the data worth evaluated by PreDIA on the y-axis and the one evaluated by EnKF on the x-axis. For measurements of hydraulic conductivity (blue points), the dominating linear dependency leads to comparable data impact estimates for the first four plots. For later data points, the deviations of the linearized estimate indicate that even hydraulic conductivity is nonlinearly dependent on other measurements. In contrast, the hydraulic head measurements (red points) clearly show the nonlinear effects for all data point places. The nonlinear dependency is not recognized by the EnKF and therefore the data impact of these designs is underestimated by the EnKF. This is in accordance with my claim that linear estimates possibly underestimate data impact for nonlinear systems.

Overall, this leads to a significantly worse performance in reducing the uncertainty associated with the plume center, even though the EnKF captures the uncertain boundary condition rea-

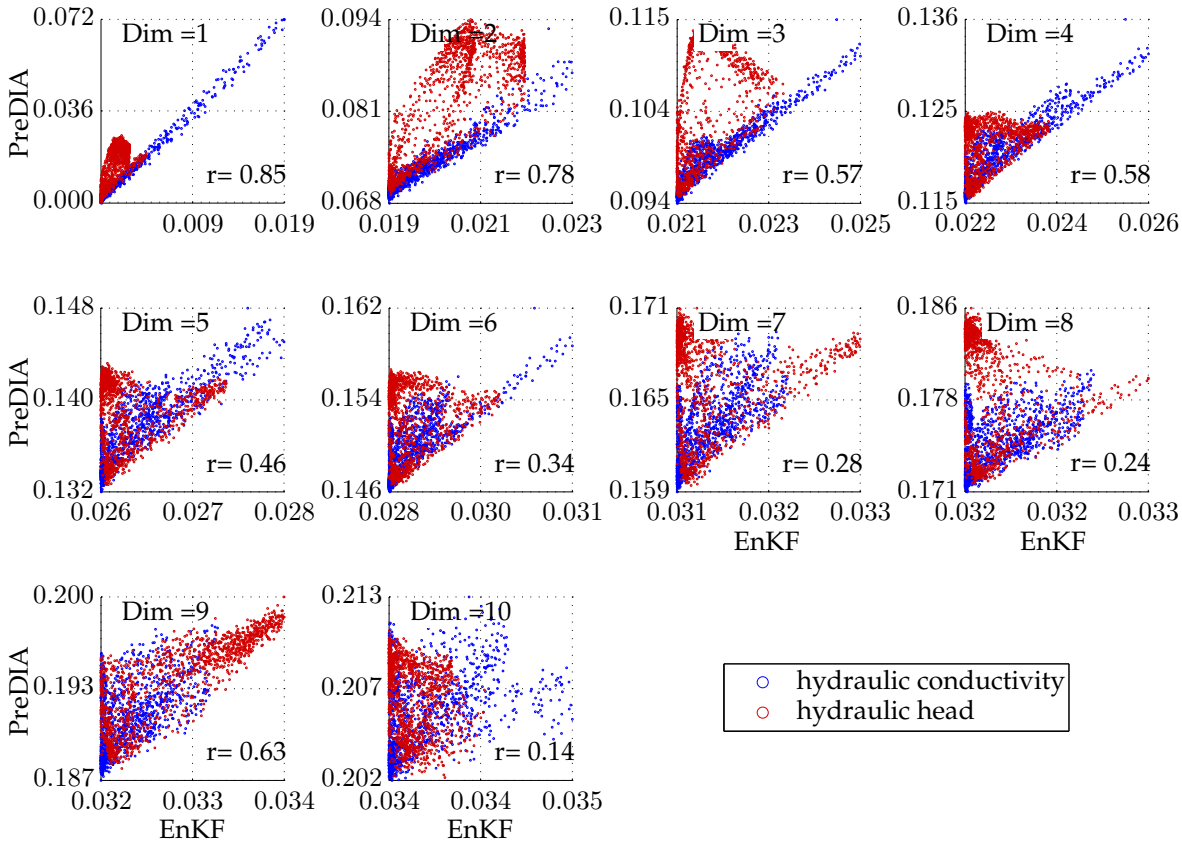


Figure 4.8.: Scatter plot comparing the data impact values obtained from data worth maps by PreDIA and by the EnKF for the design size (dimension) one to ten (TC1). Red points indicate measurements of hydraulic head, blue points indicate measurements of hydraulic conductivity.

sonably well. This can be seen by comparing the expected conditional variance within Fig. 4.6 (left and right). With a higher relative emphasis on the mostly linear source transmissivity information, the plume width and total solute flux are determined comparably well. Still, the overall prediction quality of concentration c is reduced by ignoring and misinterpreting nonlinear information, such that PreDIA clearly outmatches the EnKF. In this scenario, PreDIA achieves 25 % more uncertainty reduction with the same number of sampling positions than the EnKF.

Generalization:

In more general terms, EnKFs and all linear(ized) methods can only measure correlation, which is an incomplete assessment of statistical dependence. For example, zero correlation between a zero-mean variable and its square does not imply that a squared value is independent of its

square root. Hence, the limitations of linear(ized) methods illustrated in the specific example generalize to all nonlinear applications.

Sampling patterns optimized for predicting exceedance probability (Case 2a & 2b)

In this test case, it is the goal to maximize the prediction confidence, whether a critical concentration value (e.g. imposed by a regulatory threshold) will be exceeded or not. The PreDIA-based sampling patterns for cases 2a and 2b are shown in Fig. 4.7, again obtained from the same sample.

Case 2a ($c_{crit} = P_{15}$) exhibits a sampling pattern which is mainly based on head measurements at near-boundary and towards-target locations. Transmissivity measurements exploring the source region are practically absent. For predicting low threshold values, it is only important (and therefore sufficient) to know that the plume misses the sensitive location. This information is obtained by head measurements flanking the plume, which can reveal transverse gradients that could divert the plume from hitting the sensitive location.

Case 2b ($c_{crit} = P_{85}$) shows an inverted behavior, where the source is sampled repeatedly using six transmissivity samples that are hardly distinguishable in Fig. 4.7. Two additional transmissivity samples north of the source support the near-source samples by addressing the contrast in transmissivity between the source and its surroundings. Instead, head measurements closely flanking the plume are disregarded. This is a direct consequence of the different information needs between case 2a $\&$ 2b. For high threshold values, it is necessary to know whether the plume preserves its initial peak concentration over large travel distances up to the sensitive location. Highly conductive sources favor this behavior, and can be identified by increasing the source sampling density. In addition, highly conductive sources statistically imply an increased downstream plume width. With the plume sufficiently wide, the chances of bypassing the sensitive location by meso-scale meandering decrease and only a globally rotated mean flow direction can prevent the plume from hitting the sensitive location. That is the reason why (1) transverse gradients and the related head measurements are not closely flanking the plume, and (2) there are more remote head samples at the northern and southern boundaries that help to infer the global flow direction.

Comparison of goal oriented performance

In order to emphasize the task-specific character of the individual design patterns towards their respective prediction goal, I applied each design pattern to the prediction goals of all other test cases. This yields the performance indices summarized in Tab. 4.3.

The performance indices show that the PreDIA-based design pattern (1a) clearly outmatches the EnKF (1b) for all three prediction goals. The EnKF-based design pattern is even surpassed in its own objective by the PreDIA-based sampling patterns designed for cases 2a (low threshold) $\&$ 2b (high threshold). The worst performance was found for pattern 2a (low threshold)

	Pattern 1a	Pattern 1b	Pattern 2a	Pattern 2b
Case 1a/b	100.00 %	75.14 %	79.10 %	95.99 %
Case 2a	81.41 %	76.03 %	100.00 %	69.01 %
Case 2b	90.43 %	38.79 %	27.54 %	100.00 %

Table 4.3.: Performance indices for every sampling design when applying on different prediction goals.

when applied to the objective of case 2b (high threshold). This can be explained by the fact that these two patterns lay their focus on opposed features in their respective design objectives, i.e. on meso-scale meandering versus source conductivity. The opposite case (applying pattern 2b to case 2a) performs better. Obviously, in the specific examples, many source conductivity measurements are more generic all-purpose information than head measurements populating the boundaries.

Sampling patterns accounting for conceptual model uncertainty (Case 3)

Sampling pattern

The optimized sampling pattern for case 3 is shown in Fig. 4.9. Opposed to the previous cases, case 3 also considers conceptual model uncertainty, represented by a possibly present hydraulic barrier. If present, the barrier causes a flow regime which forces the plume to swerve northwards and so increases the chance that the plume hits the sensitive location. The strong dependence of the predicted concentration on the presence of the hydraulic barrier requires an adequate model choice. Therefore, the sampling pattern reacts to this additional uncertainty. Compared to case 1a, three transmissivity measurements are placed in the area of the possibly present barrier, while most other design features are preserved.

Support of model choice

Although model choice is not implemented as a utility function for the design (the importance of model choice is only *implicit* via its role in the chosen prediction goal), the reliability of correct model choice is improved by the adapted sampling pattern provided by PreDIA. This is a secondary affect of minimizing prediction confidence, as model choice has its own contribution to predictive uncertainty (compare e.g., Eq. (2.32)). This effect can be illustrated best by computing the preposterior weights of the two different hypothesized models: Among all possible data sets generated with the barrier, the model with barrier obtains (on average over all those data sets) a weight of 98%. Among all possible data sets generated without the barrier, the model without the barrier receives an average weight of 50%. Weighting both preposterior cases by their prior probabilities to occur (i.e. 70% and 30% respectively) yields an expected reliability of 85% to choose the correct model. This is a significantly increased reliability compared to the prior stage, where the reliability lies at 58%.

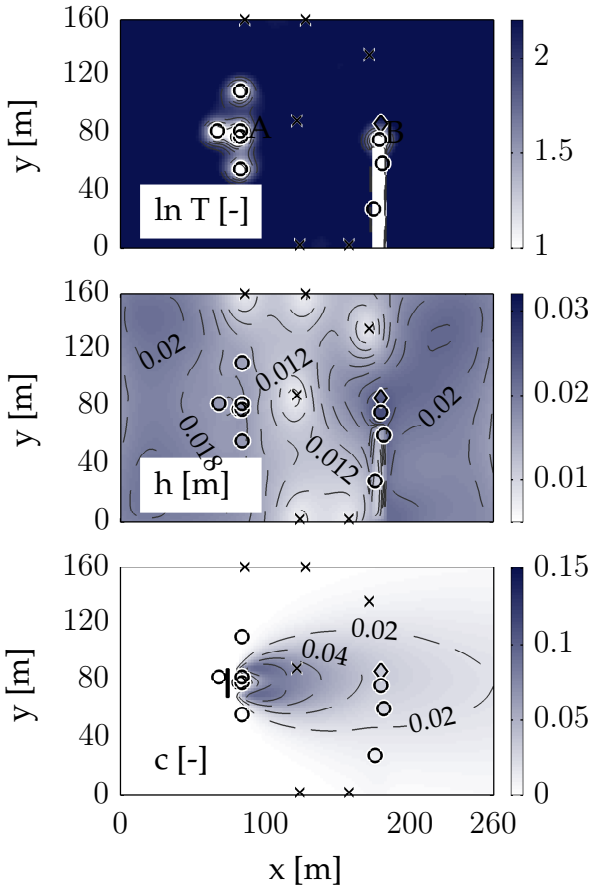


Figure 4.9.: Sampling pattern (case 3) when considering conceptual model uncertainty exemplary represented by a hydraulic barrier.

PreDIA allows one to perform such a full BMA analysis, because all these statistics are available. As for the computational costs and convergence issues, the AESS drops in scenario 3 from 500 (cases 1 and 2) to about 200. This is due to the increased variability and uncertainty in hydraulic conductivity introduced by the possibly present hydraulic barrier.

4.6. Application for a plant model scenario

This section presents an application featuring discrete BMA for soil plant model selection and was partly published in *Wöhling et al.* [2013].

Objectively selecting appropriate models for realistic simulations of coupled soil-plant processes is a challenging task since the processes are complex, not fully understood at larger scales and highly nonlinear. Also, comprehensive data sets are scarce, and measurements are uncertain. In the past decades, a variety of different models have been developed that exhibit a wide

range of complexity regarding their approximation of processes in the coupled model compartments [*Priesack and Gayler*, 2009]. Therefore, it would be extremely valuable to know even before field experiments have started, which data should be acquired at which frequency to ensure maximum confidence model selection and a minimum predictive error variance, yet at small experimental costs.

The optimal identification of the best model structure from among a suite of plausible models is an example for implicit Bayesian Model Averaging (Sec. 2.3.4) within the PreDIA method. It allows to monitor how BMA weights react to conditioning on possible data sets from different proposed sampling campaigns. In this case, I selected four different plant growth models that are coupled to a common soil water flow model. The plant growth models utilized herein are CERES [*Ritchie et al.*, 1985], SUCROS [*Goudriaan and Laar*, 1994], SPASS [*Wang*, 1997; *Gayler et al.*, 2002], and GECROS [*Chapman*, 2008]. These four disparate models are all incorporated in the model system Expert-N [*Engel and Priesack*, 1993; *Biernath et al.*, 2011]. These simultaneously describe evapotranspiration, root water and solute uptake, soil heat fluxes, and plant growth processes at different levels of detail and abstraction. *Priesack et al.* [2007] used CERES, SPASS and the SUCROS model to investigate the impact of crop growth model choice on simulated water and nitrogen balances. They found only subtle differences among the different models in their simulation of the water balance, but comparatively large differences in their performance to predict C and N turnover. More recently, *Biernath et al.* [2011] used the CERES, SPASS, SUCROS, and GECROS models to evaluate their ability to predict different environmental impacts on spring wheat grown in open-top chambers. The most adequate simulation results were obtained by SUCROS, followed by the SPASS, GECROS and CERES models respectively. It was concluded that the more mechanistic plant growth models, GECROS and SPASS, do not necessarily exhibit better predictive performance.

Using these four crop models, the main aims of this study are

- (i) to show the potential of PreDIA for the purpose of optimal identification of model structure within the BMA context,
- (ii) to investigate how BMA weights react on different available data types, such as evapotranspiration, leaf-area-index, and soil moisture data and
- (iii) to analyze the different structural deficits of the four models, by calibration on different combinations of the available data.

For a detailed description of the plant models, I refer to *Wöhling et al.* [2013] and to the above model-specific citations. Soil water flow is modeled with the Richards equation implemented in the HYDRUS-1D software [*Simunek et al.*, 2005] (see Sec. 2.1.2).

The van-Genuchten Mualem model [*Genuchten*, 1980] is used to parameterize the soil hydraulic functions. In all simulations, two horizontal soil layers are assumed with depth ranges from $0 - 0.12m$ and $0.12 - 0.21m$. Five common soil hydraulic parameters are selected for each of the two soil layers (totaling ten soil hydraulic parameters) and four specific crop model parameters as uncertain parameters in the modeling scheme. These parameters appeared most sensitive to the model predictions and little knowledge was available a priori. The uncertain soil parameters are the van-Genuchten parameters for the saturated water content, $\theta_s[M^3M^{-3}]$,

the shape parameters of the water retention function, α and n, the saturated hydraulic conductivity, $K_s[-]$) and the pore-connectivity parameter $l[-]$.

4.6.1. Test case

I consider three different data types within the data impact analysis: evapotranspiration (ETA), leaf area index (LAI), and soil moisture. Individual data packages and combinations thereof are used to condition ensembles containing random realizations of the four soil-plant models. Combining these data types in different groups results in eight different data packages, corresponding to the eight different Bayesian model update runs listed in Tab. 4.4.

Data package y_n	0	1	2	3	4	5	6	7
LAI					x	x	x	x
ETA			x	x			x	x
θ (5 cm)		x		x		x		x
θ (5 cm)		x		x		x		x

Table 4.4.: Definition of the different investigated data packages. The x marks which data types are includes within the data package 1-7, represented each as a column.

4.6.2. Results and discussions

Data impact analysis for model choice

In the very first step, I compare the performance of the four single crop growth models to the performance of the BMA mean when conditioned on all data (run 7).

To do so, I calculated the normalized root mean squared error (NRMSE) for LAI, ETA, and soil moisture data and aggregated it to a single sum-of-squared-error performance criterion. This criterion attained values of 26.3, 25.6, 24.7, and 27.2 for the CERES, SUCROS, SPASS, and GECROS models, respectively. The BMA mean slightly outperformed all individual models with NRMSE = 24.1. The posterior weight distribution of the corresponding BMA run is depicted in Fig. 4.10 (run 7, rightmost bar). Approximately 48% of the weights are associated to the CERES model (red), 30% to the SPASS model (blue) and 21% to the SUCROS model (green). The contribution of the GECROS model (yellow) in the posterior weight distribution is negligible ($< 1\%$).

Considering further the posterior weight distributions of the BMA runs with the different data packages, it can be observed that GECROS also attains very low weight contributions for most of the other runs. The only exception is run 1, where the ensemble is conditioned only on soil moisture data. Here, GECROS gathers about 31% of the weights (Fig. 4.10). These findings, however, should not be used as a justification to remove GECROS from model ensembles in general. GECROS performed superior to other models at a different experimental site with a

deep loess soil [2, 19]. Potential reasons for the poor performance at the site investigated here could be a poor parameterization of the interplay between root water uptake in the shallow soil and the evapotranspiration processes.

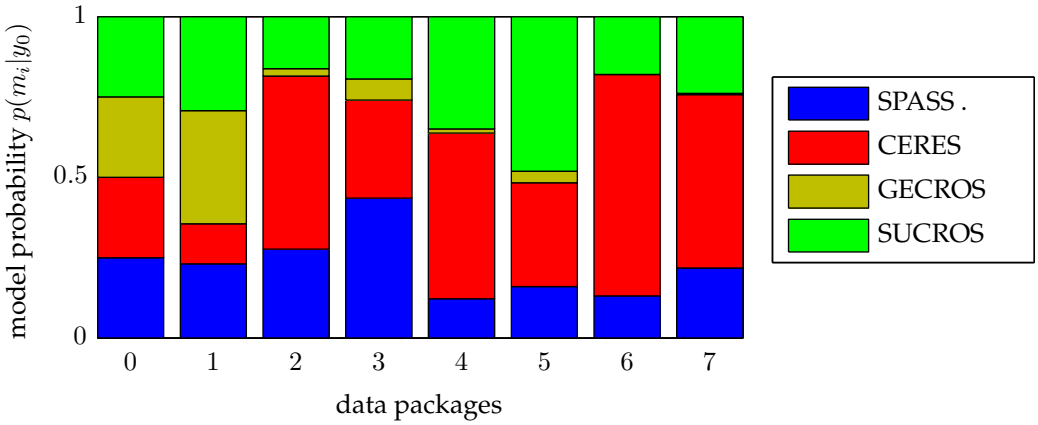


Figure 4.10.: Model weights for the four different crop models evaluated by the nonlinear inference within the PreDIA method for different data packages.

The resulting posterior distribution of BMA weights from the various runs with different data packages (runs 1 - 6) differ widely from the fully conditioned distribution (run 7). Hence, different data types (and combinations thereof) support different crop growth models. CERES obtains large weights in most cases when LAI and/or ETA is used (64% in run 6). SPASS obtains large weights for runs with ETA and θ (55% in run 3) and in cases where LAI is not present. This is an indication that SPASS performs better than the other models for processes linking soil water transport, root water uptake and evapotranspiration. This is due to its more detailed representation of root dynamics. SUCROS obtains generally smaller weights than CERES and SPASS with the exception of run 5 (44%), where LAI and θ was used. The posterior weight distribution that most closely resembles the distribution of the fully conditioned run is run 2, where only ETA was used (Tab. 4.4 and Fig. 4.10). In that sense, ETA has the largest individual data worth for finding the final model weighting achieved in this study with all data at once.

Model ensemble performance

It is challenging and expensive to measure evapotranspiration (ETA) in the field for larger areas. The installation of an eddy covariance station requires a high level of skill and experience. Also, the high-frequency data have to undergo several aggregation and post-processing steps that may introduce errors. In addition, the footprint of eddy-covariance measurements is site-specific and difficult to estimate. All this taken together can result in measurement uncertainties of $\pm 20\%$ of absolute daily values and larger. Yet it is important to correctly simulate ETA since it is a boundary flux that couples land-surface-atmosphere processes and their feedbacks.

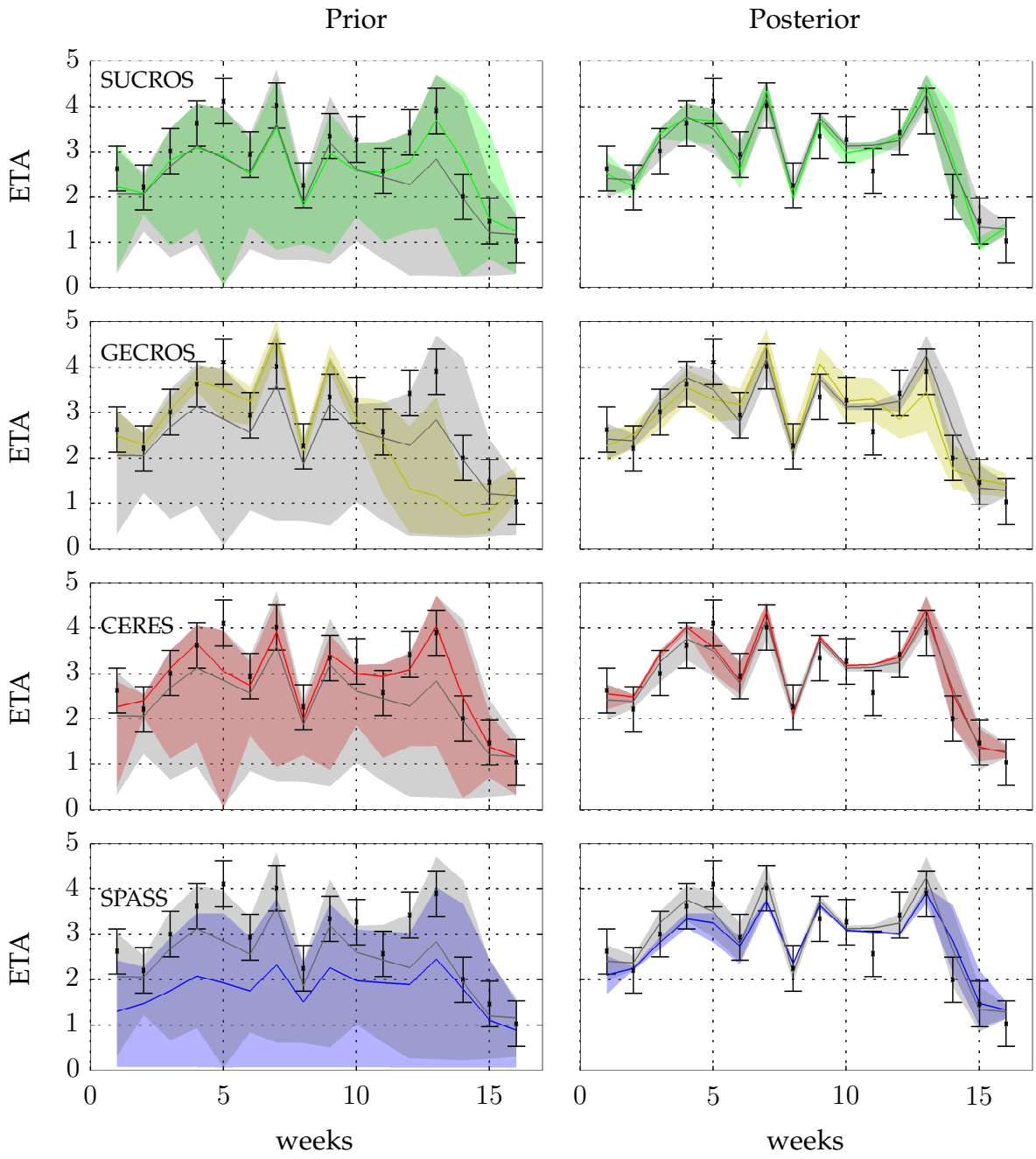


Figure 4.11.: Prior and posterior simulation of ETA: The BMA prediction is depicted in gray, whereas the individual models are printed in color. The expected value is augmented by the 0.1 and 0.9 confidence intervals. The real measurements values are plotted in black with the assumed errors.

To investigate the performance of the individual models for simulating weekly averages of daily ETA in more detail, consider Fig. 4.11 which depicts the prior and posterior confidence

intervals of ETA predictions for the individual models. The posterior distributions are calculated using all available data types (run 7). The grey shades depict the 80% prior/posterior confidence intervals of the full ensemble with all four crop growth models in BMA, whereas the colored shades indicate the confidence intervals of the individual models. The prior uncertainty bounds of the CERES and SUCROS models are very large and partly exceed the corresponding uncertainty bounds of the entire ensemble. This indicates that the ETA predictions with these models are highly uncertain prior to calibration, but the uncertainty bounds cover the corresponding ETA observations. In contrast, the prior uncertainty bounds of the more complex SPASS and GECROS models are smaller, but cover less of the observations. Interestingly, the prior uncertainty of the ETA value in week 8 (GECROS, SPASS) as well as in weeks 10 and 11 (SPASS) is very small compared to other weeks (Fig. 4.11). This is also the case for the posterior uncertainty bounds that are generally smaller as expected. Solar radiation and potential evapotranspiration was low in week 8 compared to the other time intervals which could explain the higher confidence of the models, but this was not observed in weeks 10 and 11.

The uncertainty of the individual models generally increases with plant aging. GECROS is the only model that is unable to predict the ETA peaks in weeks 12 and 13 although it exhibits increased uncertainty bounds for these dates, which explains why this model obtained little weight in the BMA run with all data.

4.7. Summary and conclusions

In this chapter, I introduced a nonlinear method for information processing, called PreDIA. It assesses the expected data utility of proposed sampling designs, such as the expected impact of data on prediction confidence, in an optimal design framework. The method operates via a purely numerical Monte-Carlo implementation of Bayes theorem and Bayesian model averaging, combined with an analytical marginalization over measurement and model errors. Since the actual measurement values at the individual planned measurement locations are unknown during the planing stage of optimal design, PreDIA averages the utility of designs over all possible measurement values that a given sampling design could produce. The method can be seen as an extension of nonlinear inference by using the Bootstrap Filter (BF). Due to its full numerical character, PreDIA allows one to incorporate various sources of uncertainty and is able to manage highly nonlinear dependencies between data, parameters and predictions with ease.

I applied the method to a DoE problem taken from contaminant hydrogeology, where I illustrated its applicability to different sources of uncertainty, various prediction tasks and task-driven utility functions. Within a groundwater quality example, I considered non-co-located hydraulic head and transmissivity measurements. In order to show the limitations of linearized methods, I compared the optimal design patterns obtained via PreDIA to those from an EnKF. The following conclusions are most important:

1. PreDIA outmatches linearized methods (such as EnKFs) because linear methods fail to recognize relevant nonlinear relations between potential measurement locations and the

prediction goal. Hence, linear methods oversample locations considered to be most informative from the limited viewpoint of linearized analysis.

2. PreDIA can handle arbitrary task-driven formulations of optimal design. I demonstrate this in a scenario variation that involves predicting the exceedance of a regulatory threshold value, which is important for risk management [e.g., *de Barros et al.*, 2009]. The sampling pattern for the task-driven prediction strongly depends on the level of the threshold value, because different information needs are triggered by the underlying flow and transport physics. Neither this difference nor such classes of task-driven formulations could be handled by linearized methods.

3. The number of Monte-Carlo realizations needed by PreDIA for convergence rises with the number of planned sampling points and their measurement accuracy. This is inherited from the applied filtering techniques in general. The averaged effective sample size (AESS) serves as an appropriate measure to monitor statistical convergence. Averaging analytically over measurement and model-structural error and over the yet unknown data values drastically improves convergence. However, the problem of filter degeneracy is still a challenge when planning extensive sampling campaigns. An extension of PreDIA towards more efficient stochastic methods would help to further increase the affordable sampling size. Here, linear methods are superior as they benefit from fast analytical solutions.

4. Bayesian model averaging is implicit in PreDIA at no additional conceptual costs, and allows to reduce the subjectivity of prior assumptions on, e.g. geostatistical parameters, boundary parameters or physical/conceptual model alternatives (like hydraulic barriers). Introducing more variability to models might increase the computational costs or lead to a decrease in the AESS.

5. The specific illustrative example showed that the uncertain direction of a regional groundwater flow has a significant impact on the uncertainty of predicting contaminations, and should hence not be neglected. This additional uncertainty can be quickly reduced by hydraulic head measurements at large distances.

6. In the specific case, the optimal design predominantly addressed uncertainty in head boundary conditions and contaminant source hydraulics, rather than structural uncertainty in the geostatistical model. This will change according to the relative importance of individual sources of uncertainty, and the availability of data types that are adequate to address these individual uncertainties.

7. The plant growth example showed the flexibility of PreDIA to be used within arbitrary model contexts and also within a BMA framework to specifically identify those data that have the highest potential to discriminate different structural models. Thereby, the individual BMA weights of each model are easily extracted from the PreDIA weighting matrix at no additional costs .

This completes **Step I**, which was the introduction of a flexible and accurate reference method for data impact estimation. It will be used in the upcoming chapters as a reference method for comparison with faster and possibly approximated estimators.

5. Reverse data impact assessment

This chapter conforms to **Step II** within this thesis. It focuses on concepts to speed up the expensive nonlinear data impact estimation that was introduced in the last chapter. I will revise the nonlinear estimation process of the last chapter with regard to an information-theoretic background. Thus, I will focus on entropy as an uncertainty measure instead of the variance. This formulates the utility function of DoE of data acquisition as a measure of information that data offer for the prediction target quantity.

Based on Bayes' Theorem, the theoretic evaluation of information is symmetric with regard to its direction. Based on this symmetry, one can measure the information that hypothetical knowledge of the prediction quantity offers on potential data and use the results as a measure for data impact. Therefore, the DoE problem is re-stated as the search for those data packages about which the prediction quantity offers the most information.

This theoretically equivalent assessment of data impact will allow dramatically faster evaluation times. I will introduce this theoretical potential in Sec. 5.2 and will investigate how much of this potential can be exploited in a practical implementation in Sec. 5.3. The implementation is finally compared with the reference methodology from the previous chapter in a synthetic study in Sec. 5.6 in terms of accuracy and evaluation speed.

5.1. Introduction

Apart from efficient search algorithms to solve the optimization problem, the computer time for a single call to the utility function is a key factor for the overall computational costs within DoE, and can pose enormous challenges even in high-performance computing [e.g., *Reed et al.*, 2013]. Therefore, I focus in this chapter on the time for evaluating the utility function or, more precisely, for evaluating the underlying data impact assessment (DIA) per trialled design.

The last chapter coupled a high-dimensional DoE problem with a fully nonlinear and non-parametric Bayesian DIA. It used an efficient Bayesian framework based on the bootstrap filter to reduce the computational time required for preposterior data impact analysis. Still, I found that the substantial advantage of nonlinear DIA compared to linear approximations is paid by a tremendous increase in evaluation time. In the current section, I will show that the preposterior analysis in combination with entropy as uncertainty measure is a direct evaluation of Mutual Information (see Sec. 2.4). This will form the basis for the speedup potential exploited in the current chapter.

Reverse mindset The reverse mindset in DIA is new. It is opposed to the intuitive nature of models as directed input-output relation, and also to the intuition that measurement data offer information about the prediction. The reverse idea is that hypothetical knowledge of the prediction quantity offers information about potential data and that this information is the very same quantity as the traditional forward-related information. Because the reverse mindset in new, I want to illustrate the general idea within a fictive example and show the differences between the forward and reverse mindsets.

For this illustration, I use a simple and empirical rain prediction model, which has solely illustrative purposes and no scientific value, but is intuitive to understand. I consider three observable input parameters ($\mathbf{y}$): the perceived temperature (T), the current cloud coverage of the sky (C) and the perceived humidity of air (H). The model prediction (z) is rain (R). All four quantities can take three values each: high ($\uparrow$), medium ($\rightarrow$) and low ($\downarrow$). Our model to forecast rain consists of ten simple rules that are provided in Tab. 5.1. They are derived from typical experiences of weather conditions. The prediction of rain is evaluated according to the matching rules per parameter realization. If multiple rules apply, the prediction is the average output of all rules, rounded to the nearest discrete value. The prediction remains fully random in case that no rule applies to a parameter set.

Temp.	Cloud coverage	Humidity of air	Rain	Rule
-	$\downarrow$	-	$\downarrow$	sunny day
-	$\uparrow$	$\uparrow$	$\uparrow$	cloudy day
$\downarrow$	-	$\downarrow$	$\downarrow$	dry cold day
$\uparrow$	$\uparrow$	-	$\uparrow$	summer storm
-	$\rightarrow$	$\uparrow$	$\uparrow$	wet autumn day
-	$\uparrow$	$\uparrow$	$\rightarrow$	foggy day
$\rightarrow$	$\rightarrow$	-	$\rightarrow$	spring day
$\uparrow$	$\downarrow$	$\downarrow$	$\downarrow$	hot day
$\downarrow$	$\uparrow$	-	$\uparrow$	raining day
-	$\uparrow$	$\downarrow$	$\rightarrow$	spring day

Table 5.1.: Ten applied weather system rules. Prediction and parameter values (high, medium, low) are indicated by $\uparrow$, $\rightarrow$ and $\downarrow$. The $-$ indicates that a data type is not used by the rule.

As prior analysis, the parameters are sampled from their prior distribution (here: independent and uniform) in a Monte-Carlo analysis. The outcome of the Monte-Carlo analysis (here: $n = 1000$ realizations) is depicted in the first row of Fig. 5.1, where the lines in the parallel-axis plot connect the input parameter values (T,C,H) and the model output values (R) for each realization. In the first row on the right side, the prior distribution of the rain prediction is shown in a histogram, showing roughly uniform probabilities.

For rating the importance of different parameters considered as potential data to collect, conventional data impact analysis evaluates which parameter one needs to observe in order to gain the most additional knowledge about future rainfall. The forward DIA requires to go through

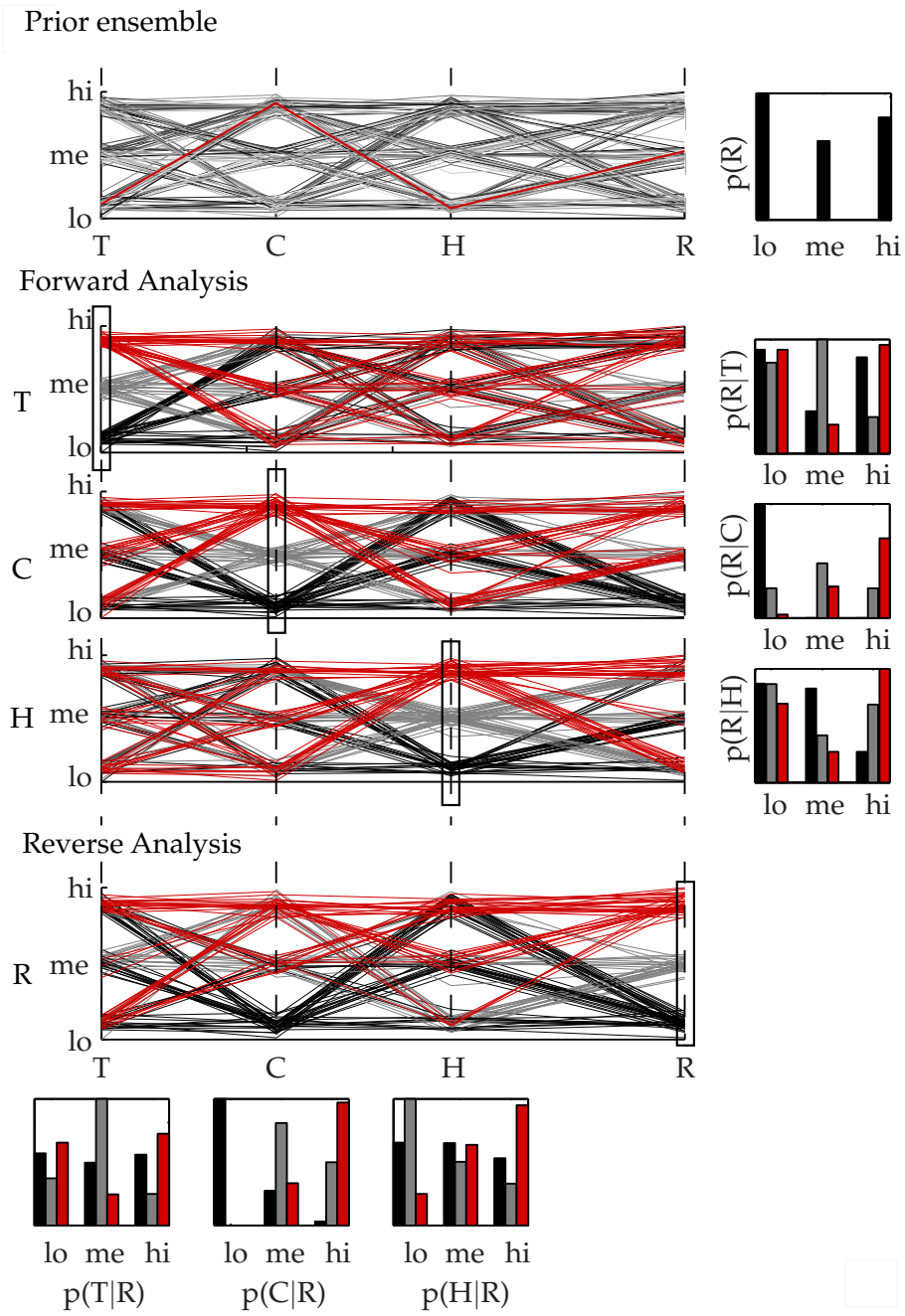


Figure 5.1.: Comparison of the forward vs. the reverse analysis in the weather illustrative example. The first row shows unconditional realizations of the model, one highlighted in red (applying rule three and nine), in a parallel axis plot. Rows two to four demonstrate the forward analysis by repeated conditioning on each possible value of each possible observation type. The conditioning is visualized as coloring all lines that pass through a possible value with one color (here: red, medium-gray, low black). The last row illustrates that one conditioning step in the reverse analysis suffices to evaluate the same dependencies. The histograms show the obtained corresponding conditional probabilities.

all possible values (high, medium, low) for each individual observable parameter. For example, when observing that cloud coverage is low, I know for sure that there will be no rain. In Fig. 5.1, this means to pick only those realizations that correspond to low cloud coverage (colored in black in the C-row in Fig. 5.1), and then to recognize that they only lead to low rain and thus to absolute certainty (see also the conditional histogram on the right side). This is an example for a high data impact. However, observing a medium cloud coverage does not allow any specific conclusion on rain (see gray lines and histogram bars in the C-row of Fig. 5.1), which is an example for a poor data impact. Observing high coverage is again more informative, as a clear trend can be observed in the corresponding conditional histogram. The process to condition on all possible values (one by one) of a considered data type to obtain a distribution of conditional distributions is called preposterior analysis [e.g., *James and Gorelick*, 1994].

The expected data impact for the data type of cloud coverage is obtained by averaging the data impact of cloud coverage over all its possible observable values multiplied by the respective occurrence probabilities. Next, one compares this result to the results from the same analysis repeated for temperature and humidity. From that comparison, one can learn that cloud coverage is the parameter with the largest individual predictive power, i.e., with the highest data impact. Overall, this requires conditioning the ensemble repeatedly for each possible data value of each possible data type.

The reverse DIA, in contrast, only requires conditioning on the possible values of the prediction (rain). In Fig. 5.1 (R-row), one only needs to color the ensemble according to the predicted rain level and then one can see whether this sorting also leads to a good color sorting for any of the observable parameters. At one glance, one can see that this is the case for cloud coverage, which is qualitatively equivalent to the findings of the forward analysis. Thus, coloring (resembling the evaluation of the conditional distribution) is only required once and becomes therefore a preprocessing step outside the loop that searches for the best data type among the three available ones. The gained speedup factor is, in this example, a factor of three. I will revisit this example in Sec. 5.5 and demonstrate the exact quantitative equivalence.

While this very simplified example illustrates the general idea, the revealed advantages of the reverse analysis become much more obvious for complex systems, which differ from the current example in the following ways:

- The list of possible data types at different locations is in the order of thousands or more. Therefore, swapping the conditioning exercise between data and model output is drastically more rewarding.
- Possible observation and prediction values are continuous rather than discrete (here: high, medium, low), making the averaging over all possible values and the conditioning computationally more expensive.
- Estimation of joint data impact for multiple measurements to be taken simultaneously requires averaging the data impact for all possible combinations of data values by their joint *pdf*. Also, the individual data impact per possible data set is obtained by a multivariate conditioning problem. In this setting, however, the reverse analysis still includes only conditioning the prediction target, which remains a univariate conditioning exercise.

These differences lead to challenges for both the forward and the reverse analysis, which require specific solutions and numerical implementation (see Sec. 5.4.2).

5.2. Reverse approach

The goal of this section is to provide a mathematical foundation for the reverse methodology in nonlinear data impact analysis. The key idea is to reverse the direction of information analysis for assessing data impact, by swapping the roles of observable data and model prediction in the analysis of statistical dependency. Please note that, in the field of linear sensitivity analysis, reverse approaches have a successful history. Examples are adjoint-state sensitivity analyses (see Sec. 3.3.1), where the adjoint states are subject to the same physics as the original state, but with a reversal of causality or time. Similar derivations led to the reverse formulation for linear transport in steady-state flow fields [e.g., *Neupauer and Wilson*, 2001].

In contrast to these existing techniques, this approach reverses the direction for analyzing nonlinear statistical dependency, rather than reversing time or causality in a differential equation for linear(ized) sensitivity analysis. Nevertheless, the expected benefits are the same, i.e., to speed up a problem solution by choosing the smallest possible dimension to solve an expensive problem.

Symmetry: For for each design candidate $\mathbf{d}$ proposed during optimization, nonlinear data impact analysis requires to simulate multiple future values of the potential measurement data sets $\mathbf{y_d}$ (see details in Chap. 4). These are used for Bayesian inference to evaluate their data impact per data realization related to a pre-defined model prediction target z. The conventional forward data impact analysis follows the natural understanding to investigate how the uncertainty in the relevant prediction quantity z would decrease when conditioning on a vector of possible future observation data $\mathbf{y}_d$. The relevant probability density functions (*pdfs*) for assessing statistical dependency are therefore the prior *pdf* $p(z)$ and the conditional *pdf* $p(z|\mathbf{y_d})$ of the prediction. A rearrangement of Bayes' Theorem, which has been introduced in Sec. 2.3.2, shows that the direction in which statistical dependency is assessed does not matter in data impact analysis, regardless of the involved system models:

$$p(z|\mathbf{y_d}) = \frac{p(\mathbf{y_d}|z)p(z)}{p(\mathbf{y_d})} \;\Rightarrow\; \frac{p(z|\mathbf{y_d})}{p(z)} = \frac{p(\mathbf{y_d}|z)}{p(\mathbf{y_d})}. \tag{5.1}$$

The *pdf*-ratio $p(z|\mathbf{y_d})/p(z)$ on the left side of Eq. (5.1) relates the posterior state $p(z|\mathbf{y_d})$ to the prior state $p(z)$ as in the conventional (forward) data impact analysis. The equivalent right side of Eq. (5.1) is the foundation of the new reverse analysis direction of nonlinear DIA, as it relates the posterior state of $p(\mathbf{y_d}|z)$ (conditioned on a model prediction) to the prior state of $p(\mathbf{y}_d)$. Thus, any utility function Φ to measure data impact that is based on the *pdf*-ratio on the left side of Eq. (5.1) can also be applied to the ratios on the right side of Eq. (5.1):

$$\Phi_{forw} \stackrel{!}{=} \underbrace{\Phi\left(\frac{p(z|\mathbf{y_d})}{p(z)}\right)}_{forward} = \underbrace{\Phi\left(\frac{p(\mathbf{y_d}|z)}{p(\mathbf{y_d})}\right)}_{reverse} \stackrel{!}{=} \Phi_{rev}, \tag{5.2}$$

and will result in equivalent estimates of data impact. The left side of Eq. (5.2) introduces the forward utility function $\Phi_{forw}(\mathbf{d})$, whereas the right side depicts the newly introduced reverse utility function $\Phi_{rev}(\mathbf{d})$. The equality in Eq. (5.2) shows that the optimization of both utility formulations will necessarily lead to the same optimal design:

$$\arg\max_{\mathbf{d}\in\mathbf{D}}\Big[\Phi_{forw}(\mathbf{d})\Big] = \arg\max_{\mathbf{d}\in\mathbf{D}}\Big[\Phi_{rev}(\mathbf{d})\Big]. \tag{5.3}$$

This equality is the foundation of the reverse approach, which analyses how the possible observation data $\mathbf{y}_d$ are influenced by hypothetical knowledge of the predicted quantity z. The approach is potentially faster than the classical forward approach for two reasons:

(1) The conditioning on possible values of the model prediction z is independent of the chosen design $\mathbf{d}$, and moves outside of the optimization loop.

(2) The model prediction of interest in the context of DoE is often low-dimensional if not even univariate (the reasoning for this is provided below), making the conditioning exercise much easier to achieve. This avoids the expensive conventional preposterior analysis that requires conditioning the model on a possibly long data vector as in Chap. 4.

These advantages will materialize in the mathematical formulation in Sec. 5.3 and in the available implementation choices in Sec. 5.4.

Applicability of the approach: One might be skeptical about the challenge to compute $p(\mathbf{y_d})$ for any given design and possible corresponding data sets $\mathbf{y_0}$, which is a well known challenge in Bayesian updating. In fact, Markov Chain Monte-Carlo methods have been developed to avoid explicit evaluation of $p(\mathbf{y_0})$. The necessity to approximate $p(\mathbf{y_d})$ for many possible data sets per trialled design is indeed one of the challenges in the suggested approach. However, the provided test cases will demonstrate that the advantages prevail.

An idea that is slightly related to the one introduced here can be found in the area of fully linear and multi-Gaussian geostatistical optimal design. Under these conditions, the ratio of prior to posterior entropy of the random field can be computed much faster by looking at the much smaller data space instead of looking at the high-dimensional parameter space [*Nowak*, 2009b; *Abellan and Noetinger*, 2010]. The reverse approach uses similar principles, but applies them to a nonlinear, non-parametric and non-multi-Gaussian analysis, looking at an even smaller (one-dimensional) model prediction space.

The reverse analysis used here focuses on evaluating the decrease in model uncertainty for the model output itself, not in terms of related monetary benefits. However, all monetary-based utility or decision frameworks relate to the model output more or less directly. As *Feyen and Gorelick* [2005] stated, it is often sufficient to evaluate the monetary cost-efficiency

only once for the optimal design, as inferior designs perform even worse. The reasoning behind this statement is that the models with the smallest output uncertainty also lead to the smallest engineering or management costs for providing safety margins. By agreeing on that, monetary-based utility functions can be excluded from the DoE framework and can be applied later only once for the optimal design.

As mentioned above, I proceed from the assumption that the dimension of the prediction space is relatively small or unity, i.e. the featured model prediction z is a single number. Of course, the model output of a transient contaminant transport simulation might be a 4-dimensional contaminant plume in time and space. However, decision-relevant quantities tend to be global aggregates of the model [e.g., *Saltelli et al.*, 2008], such as total contaminant mass fluxes, maximum concentration levels at selected compliance points or planes, peak pressures endangering a system, peak flood levels, and so forth. Thus, a model-based decision framework would merely use a 4-dimensional output as intermediate information to extract the relevant information z. In many cases, several such scalar model outputs are decision relevant. In such cases, multi-objective optimization frameworks can be applied for DoE, and the approach presented here can serve to speed up the evaluation of each individual objective:

$$\Phi_i = f\left(\frac{p(\mathbf{y_d}|z_i)}{p(\mathbf{y_d})}\right) \ , \ i = 1, .., n \tag{5.4}$$

Apart from that, the prediction z can be any aggregated quantity such as the average or maximal concentration in a area of concern, the highest water level in a river stage time series or the longest time periods below a critical water level in irrigation scenario.

5.3. Methodology: Forward-reverse equivalence of mutual information and entropy

In Sec. 5.2, the general reversibility of the theory has been derived with a yet unspecified measure Φ for data impact. The only specification was that Φ has to be a function acting on the *pdf*-ratio that appears in Eq. (5.2). In this section, I will demonstrate that Mutual Information (see Sec. 2.4) is a suitable measure that fulfills this requirement and is therefore symmetric in the direction of information processing [e.g., *Cover and Thomas*, 2006, p.251].

In this context, Eq. (2.54) can serve as forward utility function Φ_{forw}:

$$\mathrm{MI}(\mathbf{z};\mathbf{y}) = E_{\mathbf{y}}\left[h_{rel}\left(\frac{\mathbf{z}|\mathbf{y}}{\mathbf{z}}\right)\right] = \int_{\mathbf{y}} p(\mathbf{y}) \int_{\mathbf{z}} p(\mathbf{z}|\mathbf{y}) \log\left[\frac{p(\mathbf{z}|\mathbf{y})}{p(\mathbf{z})}\right] \, d\mathbf{z} \, d\mathbf{y}. \tag{5.5}$$

To obtain the reverse formulation of Eq. (2.54), Bayes' Theorem can be used to exchange $p(z|\mathbf{y_d})/p(z)$ with $p(\mathbf{y_d}|z)/p(\mathbf{y_d})$ (see Eq. (5.1)). At the same time, the joint *pdf* $p(z, \mathbf{y_d})$ expressed in Eq. (2.54) as $p(z, \mathbf{y_d}) = p(z|\mathbf{y_d})p(\mathbf{y_d})$ is replaced by $p(z, \mathbf{y_d}) = p(\mathbf{y_d}|z)p(z)$. Finally, flipping the order of integration leads to the following equality:

$$\begin{aligned}
\underbrace{E_{\mathbf{y_d}}\left[h_{rel}\left(\frac{z|\mathbf{y_d}}{z}\right)\right]}_{\phi_{forw}} &= \int_{\mathbf{y_d}} p(\mathbf{y_d}) \int_z p(z|\mathbf{y_d}) \log\left[\frac{p(z|\mathbf{y_d})}{p(z)}\right] \, dz \, d\mathbf{y_d} \\
&= \int_z p(z) \int_{\mathbf{y_d}} p(\mathbf{y_d}|z) \log\left[\frac{p(\mathbf{y_d}|z)}{p(\mathbf{y_d})}\right] \, d\mathbf{y_d} \, dz \\
&= \underbrace{E_z\left[h_{rel}\left(\frac{\mathbf{y_d}|z}{\mathbf{y_d}}\right)\right]}_{\phi_{rev}}.
\end{aligned} \tag{5.6}$$

This equation states the forward-reverse equivalence of MI. It corresponds directly to the well-known symmetry of MI [e.g., *Cover and Thomas*, 2006] in its input arguments (here: z and $\mathbf{y_d}$), which is a well-known property of MI in the field of information theory:

$$\begin{aligned}
\mathrm{MI}(z;\mathbf{y_d}) \equiv \; & E_{\mathbf{y_d}}\left[h_{rel}\left(\frac{z|\mathbf{y_d}}{z}\right)\right] \\
= \; & E_z\left[h_{rel}\left(\frac{\mathbf{y_d}|z}{\mathbf{y_d}}\right)\right] \equiv \mathrm{MI}(\mathbf{y_d};z)
\end{aligned} \tag{5.7}$$

This symmetry basically states that the random variable z offers about the random variables $\mathbf{y_d}$ the same information that $\mathbf{y_d}$ offers about z. This allows choosing between conditioning the model prediction z on possible data $\mathbf{y_d}$ or conditioning $\mathbf{y_d}$ on possible values of z within the evaluation of MI.

Relation to entropy-based approaches: Furthermore, is possible to show that other utility functions based on entropies are reversible as well. MI can be re-written as:

$$\begin{aligned}
\mathrm{MI}(z;\mathbf{y_d}) = \; & h(z) - h(z|\mathbf{y_d}) \\
= \; & h(\mathbf{y_d}) - h(\mathbf{y_d}|z)
\end{aligned} \tag{5.8}$$

where $h(\cdot)$ generally denotes the entropy of a (set of) random variables. As $h(z)$ remains constant during the optimization of $\mathbf{d}$, the following equalities can be applied:

$$\begin{aligned}
\arg\min_{\mathbf{d}\in\mathbf{D}}\left[h(z|\mathbf{y_d})\right] = \; & \arg\max_{\mathbf{d}\in\mathbf{D}}\left[\mathrm{MI}(z;\mathbf{y_d})\right] \\
= \; & \arg\max_{\mathbf{d}\in\mathbf{D}}\left[\mathrm{MI}(\mathbf{y_d};z)\right]
\end{aligned} \tag{5.9}$$

and it follows that minimizing $h(z|\mathbf{y_d})$ leads to identical designs as when maximizing MI. This includes, e.g., the well-known D-criterion for optimal design [*Box*, 1982], for which the reversibility has unintentionally been shown through linear algebra relations in linear design problems by *Nowak* [2009a] and later also by *Abellan and Noetinger* [2010]. Please note that maximizing $h(\mathbf{y_d}|z)$ does not yield equivalent designs, because the prior entropy $h(\mathbf{y_d})$ varies for different designs and, therefore, (Eq. (5.9)) would not apply.

5.4. Numerical implementation

This section offers a detailed description of possible numerical implementations for the forward formulation of MI (see Sec. 5.4.2) and for the reverse formulation of MI (seeSec. 5.4.3). The forward implementation is based on the first line of Eq. (5.6), while the reverse analysis is based on the second line. The two versions are theoretically equivalent (Sec. 5.3), but use different evaluation orders and implementation strategies that lead to the expected difference in evaluation speed.

5.4.1. Preprocessing of the ensemble

In the following, I will provide the individual techniques for conditioning, density estimation and preposterior expectation (all involved in Eq. (5.6)) for both approaches. The common starting point for both analyses is a sufficiently large ensemble $\{\boldsymbol{\theta}_i, \boldsymbol{\xi}_i, k_i, \mathbf{s}_i, \mathbf{z}_i, \mathbf{y}_i\}$ that serves to represent all required multi-variate *pdf*s. This ensemble is generated by MC simulation with random input parameter sets $\{\boldsymbol{\theta}_i, \boldsymbol{\xi}_i, k_i, \mathbf{s}_i\}$, a prediction model $z_i = f_z(\mathbf{s_i}, \boldsymbol{\theta}_i, \boldsymbol{\xi}_i) + \epsilon_z$, a data model $y_i = f_y(\mathbf{s_i}, \boldsymbol{\theta}_i, \boldsymbol{\xi}_i) + \epsilon_z$ and distribution assumptions for the data and model errors as represented by ϵ_z and ϵ_y. Because both the forward and the reverse formulation require only the joint distributions $p(\mathbf{y}, z)$, only the ensemble $\{\mathbf{y}_i, z_i\}$ will be stored.

The structural parameters $\boldsymbol{\theta}, \boldsymbol{\xi}, k_i$ are not directly part of the utility function and therefore are hidden parameters in the data utility assessment. However, the aspect of model structural uncertainty is an explicit part of generating the ensemble $\{\mathbf{s}_i, z_i, \mathbf{y}_i\}$. Thus, the obtained ensemble $\{\mathbf{y}_i, z_i\}$ represents the corresponding Bayesian distribution [*Kitanidis*, 1986] obtained by marginalizing over $p(\boldsymbol{\theta}, \boldsymbol{\xi}, k)$ (see Sec. 2.3.4).

All required realizations $\{\mathbf{s}_i, \boldsymbol{\theta}_i, k_i, z_i, \mathbf{y}_i\}$ are generated before the entire analysis, and are recycled for all steps including the outer optimization loop. Generating the ensemble is thus a one-time off-line computational effort that does not influence the speed of the optimization or data impact analysis. In the following, all steps are demonstrated for a given design $\mathbf{d}$ as it may appear within the optimization procedure.

5.4.2. Forward analysis using mutual information

Conditioning: Evaluating any posterior state of the model prediction $p(z|\mathbf{y_d})$ in the first line of Eq. (5.6) requires numerical conditioning of the model prediction z on a given measurement vector $\mathbf{y_d}$. The conditioning step is equivalent to the PreDIA method from the previous chapter (Sec. 4.2), which is based on the bootstrap filter. For Gaussian error models, computing many (high-) dimensional likelihoods requires the repeated evaluation and summation of (high-) dimensional Gaussian likelihood functions, based on the measurement error ϵ_y. This task can be outsourced using three different code libraries, which are introduced in Appendix A. The choice between these libraries will be specified in the individual test cases introduced in Sec. 5.6.

Density Estimation: The density $p(z|\mathbf{y_d})$ required in the first line of Eq. (5.6) is estimated using kernel density estimation techniques, which were introduced in Sec. 2.2.3. The weighting matrix $\mathbf{W}$ as used in the PreDIA framework (see Sec. 4.3.2) contains the conditioning information and is used to weight the realization for the conditional density estimation. Similarly as in the conditioning, an error model for the prediction quantity z needs to be applied consistently in both the forward and reverse approach. As discussed in Sec. 2.2.3, the standard deviation of the measurement error ϵ_z is used as the kernel bandwidth in the KDE to directly estimate conditional prediction *pdfs* $\hat{p}(z_e|\mathbf{y_{d},}_j)$ with $z_e = z + \mathcal{N}(0, \epsilon_z)$. The approximated density is only evaluated at a set of discrete target points z_t for the numerical integration in the entropy estimation (see below).

Entropy: The conditional entropy $\hat{h}(z_e|\mathbf{y_{d},}_j)$ for the individual realizations of measurement data sets $\mathbf{y_{d},}_j$ is estimated via the integral

$$\hat{h}(z_e|\mathbf{y_{d},}_j) = \int_{z_e} \hat{p}(z_e|\mathbf{y_{d},}_j) \log\left[\hat{p}(z_e|\mathbf{y_{d},}_j)\right] dz_e \quad ; \quad \mathbf{y_{d},}_j \sim p(\mathbf{y_d}). \tag{5.10}$$

This conforms with the inner integral in the first line of Eq. (5.6). The integral is approximated numerically with the trapezoidal rule, which is sufficiently accurate in one dimension. Therefore, equally spaced integration points z_t are used over the physical range of z_e. Other examples of one-dimensional integration rules (Gauss quadrature, trapezoidal rule in the rank space of z_e) would also be possible. The prior entropy $\hat{h}(z_e)$ is estimated similarly, using the prior distribution $\hat{p}(z_e)$ represented by the ensemble $\{z_i\}$ together with the prediction error model.

Preposterior expectation: The final step is the preposterior expectation that conforms to the outer integral $\int_{\mathbf{y}} [\cdot]\, p(\mathbf{y_d}) d\mathbf{y}$ in the first line of Eq. (5.6). Each preposterior state $j = 1...m$ results in a different conditional entropy for the corresponding realization $\mathbf{y_{d},}_j$ of the potential data sets. MC integration [*Lepage*, 1980] is used to approximate the expectation over $p(\mathbf{y})$:

$$\int_{\mathbf{y}} \hat{h}(z_e|\mathbf{y_{d},}_j)\, \hat{p}(\mathbf{y_{d},}_j) d\mathbf{y} = \frac{1}{m} \sum_{j=1}^{m} \hat{h}_j(z_e|\mathbf{y_{d},}_j) \; ; \; \mathbf{y_{d},}_j \sim p(\mathbf{y_d}) \tag{5.11}$$

where $\mathbf{y}_{d,j}$ is drawn from the prior distribution $p(\mathbf{y}_d)$. For an accurate preposterior expectation, m needs to be sufficiently large to achieve an accurate sample representation of $p(\mathbf{y}_d)$. The final estimate of the MI is the expected entropy reduction from the prior to the preposterior:

$$\phi_{forw}(\mathbf{d}) = \hat{h}(z_e) - \frac{1}{m} \sum_{j=1}^{m} \hat{h}_j(z_e|\mathbf{y_{d},}_j). \tag{5.12}$$

This is the forward formulation of MI and can be used as a utility function within DoE.

5.4.3. Reverse analysis using mutual information

Conditioning: The second Line of Eq. (5.6) is asking for conditional distributions $p(\mathbf{y_d}|z)$ of possibly observable data sets $\mathbf{y_d}$ under the proposed design $\mathbf{d}$, conditional on hypothetical

observed values of the predicted quantities z. The conditioning step simplifies tremendously compared to the forward problem, because the model prediction z is a one-dimensional quantity. Thus, high-dimensional filtering techniques such as the BF are not necessary. Instead, conditioning is performed by splitting the ensemble into q subsets $Z_j = \{Z_1 \ , .., \ Z_q\}$, which corresponds to histogram-like bins with equal rank spacing, i.e., with an equal number of realizations per subset and hence with equal statistical probability $P_j(z_e \in Z_j) = 1/q \quad \forall \ j$. Random values of the model/prediction error ϵ_z are added before the classification $z_e = z + \epsilon_z$, to consistently account for the same measurement error in the predicted quantity as in the forward analysis:

$$p(\mathbf{y_d}|z_e \in Z_j) \approx p(\mathbf{y_d}|z + \epsilon_z in Z_j). \tag{5.13}$$

Equally probable classes ensure most accurate results for both density and entropy estimation. The subsets Z_j are defined based on the prior *pdf* $p(z_e)$, which does not change during the optimization. The subsets Z_j are thus independent of the currently investigated design $\mathbf{d}$. This allows an efficient implementation in which the subsets are only defined once as a preprocessing step.

Density Estimation: The challenge in the reverse analysis is to estimate the conditional joint *pdf* $p(\mathbf{y_d}|z_e \in Z_j)$ for each subset Z_j. Density estimation can be based on different techniques (see Sec. 2.2.3). Their choice mainly affects the speed of the reverse formulation, since high-dimensional density estimation requires the most computation time in the reverse analysis. However, the relatively low required number of subsets Z_j, conditioning by classification and the resulting absence of weighting substantially speed up the procedure.

Entropy: Evaluating the conditional entropy $\hat{h}(\mathbf{y_d}|z_e \in Z_j)$ in the reverse analysis (see second line of Eq. (5.6)) requires a high-dimensional integration over $p(\mathbf{y_d}|z_e \in Z_j)$. The inner integral and its MC approximation is then:

$$\begin{aligned} \hat{h}(\mathbf{y_d}|z_e \in Z_j) &= \int_{\mathbf{y_d}} \hat{p}(\mathbf{y_d}|z_e) \log \hat{p}(\mathbf{y_d}|z_e) \, \mathrm{d}\, \mathbf{y_d} \\ &\approx \frac{1}{n/q} \sum_{i=1}^{n/q} \log(p(\mathbf{y_{d},}_i)) \quad ; \quad \mathbf{y_{d},}_i \sim p(\mathbf{y_d}|z_e \in Z_j), \end{aligned} \tag{5.14}$$

where n/q is the number of Monte-Carlo realizations in the subset Z_j, and the numerical integration points $\mathbf{y_{d},}_i$ are samples drawn from the conditional sample or subset $\mathbf{y_d}|z_e \in Z_j$. This leads to the different conditional entropy estimates for each of the classes $Z_{1...j}$. The unconditional entropy of $\hat{h}(\mathbf{y}_d)$ is calculated in the same fashion using the full (unclassified) ensemble of z_e.

Pre-posterior expectation: The final step is the preposterior expectation that conforms to the outer one-dimensional integral $\int_{z_e}[\cdot] \ p(z_e) dz_e$ in the second line of Eq. (5.6). As the *pdf* of z_e is discretized using the same classes as defined in the conditioning step with the class probabilities $p(Z_j) = 1/q$, the integral over z_e is trivially approximated as:

$$\begin{aligned}\int_{z_e} h(\mathbf{y_d}|z_e \in Z_j)\; p(Z_j) dz_e \approx &\quad \sum_{j=1}^{q} P_j\; \hat{h}_j(\mathbf{y_d}|z_e \in Z_j) \\ = &\quad \frac{1}{q}\sum_{j=1}^{q} \hat{h}_j(\mathbf{y_d}|z_e \in Z_j).\end{aligned} \tag{5.15}$$

Altogether, this yields the final utility measure based on the reverse analysis:

$$\begin{aligned}\Phi_{rev}^{MI}(\mathbf{d}) &= h(\mathbf{y_d}) - \int_{z_e} h(\mathbf{y_d}|z_e \in Z_j) dz_e, \\ &\approx \hat{h}(\mathbf{y_d}) - \frac{1}{q}\sum_{j=1}^{q} \frac{1}{n/m} \sum_{t=1}^{n/m} \log \hat{p}(\mathbf{y}_t(\mathbf{d})|z_e \in Z_j) \quad ; \quad \mathbf{y}_t \sim p(\mathbf{y_d}|z_e \in Z_j).\end{aligned} \tag{5.16}$$

5.5. Application to the rainfall example

In this section, I will briefly re-visit the illustrative example of rainfall prediction from Sec. 5.1. I will numerically evaluate the MI between individual observable parameters and the prediction, using both analysis directions.

		Cloud coverage			
	$P(R;C)$	lo	me	hi	$P_m(R)$
Rain	lo	0.316	0.086	0.010	0.412
	me	0	0.158	0.097	0.255
	hi	0	0.097	0.247	0.333
	$P_m(C)$	0.316	0.369	0.354	1.000

Table 5.2.: Joint probability distribution $P(C, R)$ with both marginal distributions $P_m(R)$ and $P_m(C)$ for the rainfall example

The joint and marginal probabilities given in Tab. 5.2 allow for the calculation of the respective conditional distributions $P(R|C)$ and $P(C|R)$. The discrete character of the rainfall example entirely avoids both the conditioning and the density estimation steps that would be are necessary for continuous variables. In the forward analysis, the discrete entropy $H(R|C) = -\sum_R P(R|C)\log(P(R|C))$ is evaluated by a summation over all discrete states of C. The resulting Kullback-Leibler divergences or relative entropy $H_{rel}(R;C) = (H(R) - H(R|C))/H(R)$ equal to $[1,\ 0.025,\ 0.345]$ for the three values of C, respectively. This numbers illustrate how data worth depends on the yet unknown measurements values in nonlinear problems. Here, the information of low cloud coverage reduces the prediction uncertainty of rain to zero, whereas a medium cloud coverage only reduces the rain prediction uncertainty by about 2.5%. The preposterior averaging by the probabilities of rain $P(R)$ provides the final forward data impact $MI_{rel}(R;C) = \sum_C H_{rel}(R;C)P(C) = 0.4468$.

The respective reverse analysis results in the relative entropy values of $h_{rel}(C;R) = [0.425,\ 0.384,\ 0.473]$ for the three possible states of rain. Despite of the different relative entropy values, the *expected* data impact $MI_{rel}(C;R) = \sum_R \mathrm{MI}_{rel}(C;R)P(R) = 0.4468$ is exactly equal to the result in the forward analysis.

5.6. Application to a groundwater contaminant scenario

To illustrate my theoretical findings, to test the accuracy of the chosen implementation and to assess the promised speedup of the reverse framework, I perform three numerical test cases from the field of hydro(geo)logy.

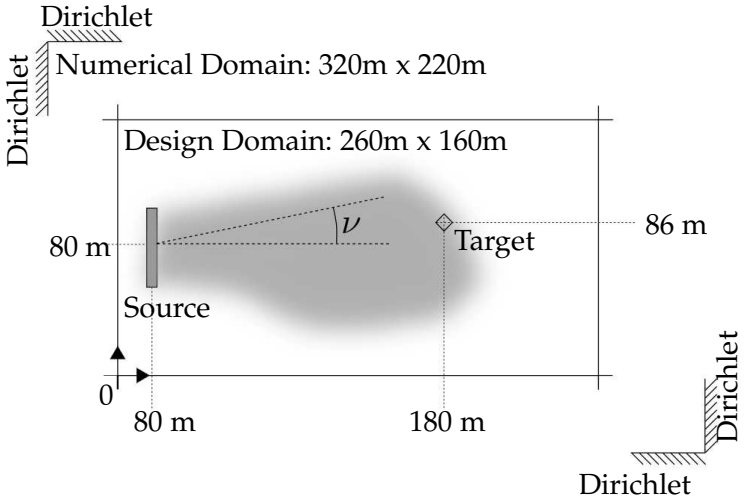


Figure 5.2.: Hydrological numerical example, including the location of the boundary conditions, contaminant source location and the target location for the concentration prediction. The design area contains all possible measurement locations that are discretized on a regular grid with two meter spacing.

The featured system in the three hydrogeological test cases is a depth-averaged heterogeneous groundwater aquifer with a long-term contaminant spill at a known location. The source width is 20 m. The system geometry is sketched in Fig. 5.2. The contaminant spreads by advective-dispersive transport. Downstream of the spill there is a point of compliance (e.g., a drinking water well or an ecologically sensitive location). At this point, different model tasks could be defined. To keep the example simple, I assume that the task is to predict the contaminant concentration at that location in the large-time (steady-state) limit.

Since a-priori prediction quality is not sufficient, it is required to collect new data. Possible data types considered here are hydraulic pressure and the hydraulic conductivity. Measurement errors for both measurement types and also for concentration values at the point of compliance are given in Tab. 5.3. The latter measurement error would apply when taking measurements of concentration in the future to validate the model prediction. Alternatively, one could interpret this as prediction model error.

The flow and transport realizations are generated with the same tools and code as described in *Nowak et al.* [2008] and *Schwede et al.* [2008]. The detailed specification of the scenario and known parameters are listed in Tab. 5.3. The parameters and their assumed prior distributions are listed in Tab. 5.4. It contains the three structural parameters $\boldsymbol{\theta}$ to define a geostatistical model for the conductivity field. The fourth uncertain structural parameter is the deviation of the mean ambient groundwater water flow direction from the x_1 direction, defined by the angle ν. The Dirichlet boundary conditions on all sides are set to match the corresponding randomized ambient flow direction in each realization. The numerical domain is extended in each direction by two average correlation lengths to avoid artificial influence of the boundaries on the statistical distribution of potential measurement values.

Numerical domain			
Domain size	$[L_1, L_2]$	$[m]$	[320, 220]
Grid spacing	$[\Delta_1, \Delta_2]$	$[m]$	[0.25, 0.125]
Design domain			
Domain size	$[L_1, L_2]$	$[m]$	[260, 160]
Grid spacing	$[\Delta_1, \Delta_2]$	$[m]$	[2, 2]
Transport parameters			
Head gradient	γ	[-]	0.01
Effective porosity	n_e	[-]	0.35
Local-scale dispersivities	$[\alpha_l, \alpha_t]$	$[m]$	[0.5, 0.125]
Diffusion coefficient	D_m	$[m^2/s]$	10^{-9}
Transversal source dimension	l_s	$[m]$	20
Known geostatistical model parameters			
Global mean	$\beta_1 = \ln T$	[-]	$\ln 10^{-5}$
Trend in mean	β_2	[-]	0
Measurement error standard deviations			
Hydraulic conductivity	$\sigma_{r,T}$	[-]	1.00
Hydraulic head	$\sigma_{r,h}$	[m]	0.01
Contaminant concentration	$\sigma_{r,z}$	[-]	0.07

Table 5.3.: Known parameters for the flow, transport and geostatistical model in the hydrological example

Geostatistical parameters			
Variance	σ_T^2	[-]	$\mathcal{N}(\mu = 4.0, \sigma = 0.3)$
Integral scale	λ	[m]	$\mathcal{N}(\mu = 15, \sigma = 2.0)$
Matérn Kappa	κ	[-]	$\mathcal{U}(a = 5, b = 36)$
Boundary parameters			
Ambient flow deviation	ν	[°]	$\mathcal{N}(\mu = 0.0, \sigma = 15)$

Table 5.4.: Uncertain structural and boundary parameters and their assigned prior distributions for the hydrological example

Test case 1 (TC1): point-wise equivalence The aim of the first test case is to demonstrate in the hydrological application example the equivalence of the numerically evaluated data impact for both analysis directions. For this comparison, an excessively large ensemble of $n = 5 \times 10^5$ realizations is chosen to minimize the statistical noise in both implementations. The forward implementation requires to approximate the expectation over $p(\mathbf{y_d})$, which is done by a number of $m = 5 \times 10^3$ preposterior states. In the reverse implementation, $q = 10$ subsets Z_j sufficiently approximate the expectation over the distribution $p(z)$ and guarantees accurate preposterior statistics. Both m and q were assessed in preliminary convergence tests, whereas higher values for m lead to better approximation of $p(\mathbf{y}_d)$, but q cannot be increased alike. High values of q lead to a better approximation of $p(z)$, but since the ensemble is split q-times, the conditional distributions are approximated with less accuracy. Thus, this trait-off between approximating $p(\mathbf{y}_d)$ and $p(\mathbf{y}_d|z)$ lead to the best combined approximation for relatively low value of m. Using the greedy algorithm (see Sec. 3.2.1) allows to compare both implementations for individual design propositions with rising dimensionality. The equivalence is illustrated as a sequence of data impact maps resulting from both formulations, Φ_{forw} and Φ_{rev} and their point-wise comparison.

Test case 2 (TC2): global optimization The second test case compares the optimal designs achieved from both utility functions, Φ_{forw} and Φ_{rev}. The optima are found in a global optimization problem using the genetic algorithm as introduced in Sec. 3.2.2. The task is to find the most informative design containing ten measurements. The focus thus shifts to the equality of both implementations for near-optimal design solutions. This analysis is repeated 20 times for both implementations to average over any stochastic noise and over the varying performance of the genetic algorithm that does not guarantee to find the global optimum. To keep the evaluation times within acceptable limits, each run is based on an reduced ensemble with $n = 1 \times 10^5$ realizations. The best, worst and average design utility out of all optimizations are then compared between the two analyses. For a fair comparison of computational time, the same number of utility function evaluations were used in all optimization runs, by fixing the number of evaluated generations to 500 for the GA.

For comparing the quality, the resulting 20 near-optimal designs were re-evaluated using the forward analysis with an ensemble size of 5×10^5 realizations. This single costly evaluation of data impact for the 20 resulting optimal design serves as a reference value to benchmark the convergence of the involved MC simulations. This reference value is used as a performance measure for comparing the data impact values of the different formulations.

Test case 3 (TC3): Evaluation time To assess the success of the reverse approach, the evaluation times are compared. Therefore, the fastest forward approach is used and competes with different implementations of the reverse approach. The different implementations use different conceptual approaches to estimate the conditional entropy. The evaluation times are compared for design taken from TC2 with differing number of measurements and ensemble size.

5.6.1. Results and discussion

The measures of uncertainty and information in this section are entropy [nats] and Mutual Information [nats], which are based on the natural logarithm. To provide a more intuitive measure for data impact, all following results will be provided as the relative uncertainty reduction, i.e., as Mutual Information MI_{rel} [-] normalized by the prior entropy of the prediction variable:

$$MI_{rel} = \frac{MI(\mathbf{y}; z)}{h_z}. \tag{5.17}$$

TC1: Point-wise equivalence

Fig. 5.3 (left) shows the relevant part of data impact maps from both analysis directions for the third measurement placed by the greedy search algorithm. The top row shows the data impact maps obtained with the forward analysis and the lower row shows the results of the reverse analysis. The two sampling locations already fixed by greedy search are indicated in black. The data worth map for the hydraulic conductivity shows the highest expected data impact in the vicinity of the source region, although the source region has already been sampled. This repeated sampling reflects the importance of sampling the contaminant source conditions for far-field predictions of contaminant transport, which has been derived and indicated by *de Barros and Nowak* [2010] and experimentally validated by *Gueting and Englert* [2013].

However, a hydraulic head measurement placed at the lower boundary (marked by the red cross) has the overall highest data impact. This measurement addresses the uncertainty in the angle of the regional groundwater flow direction (see Tab. 5.4), and helps to infer whether the contaminant plume will bypass or hit the featured point of compliance (indicated by the white diamond).

The right side of Fig. 5.3 shows the data impact maps for the eighth placement sequence within the greedy search. All four maps reveal an increased level of noise. This noise is caused by the degeneration of the bootstrap filter in the forward case (upper row) and by the high-dimensional *pdf* estimation in the reverse case (lower row). Furthermore, one can observe a bias against each other in the overall values of data impact between the two implementations (note the different color scaling), increasing with the extent of the considered sampling campaign.

The main reason for this increasing bias is the different behavior of the two implementations, when the ensemble size becomes insufficient given the growing dimensionality of the underlying filtering or *pdf* estimation task. The forward implementation has a tendency to overestimate the data impact, because too few realizations remain after conditioning due to filter degeneration, representing an artificially low level of uncertainty. In the reverse case, the high-dimensional *pdf* estimation with a given number of realizations leads to a phenomenon known as concentration of measure [*Ledoux*, 2001]: The space is sampled too sparsely, so that no density differences can be determined numerically. This phenomenon leads to an artificial convergence of the conditional distribution towards the uniform distribution, which has the highest possible

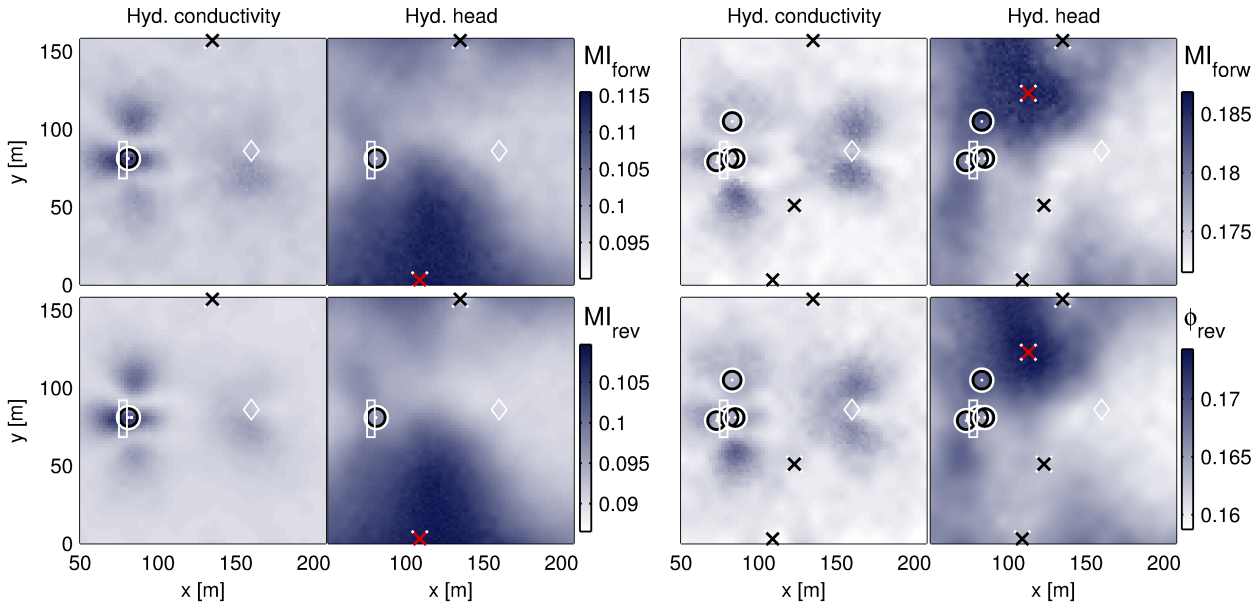


Figure 5.3.: Comparison of data impact maps for the forward (upper row) and the reverse (lower row) analysis, measured as normalized Mutual Information (MI). The plots on the left show the data impact maps for placing the third measurement in the greedy search. The maps reveal the highest data impact for a measurement of hydraulic head at the location indicated by the red cross. The corresponding plots on the right side show the data impact maps for the eighth measurement placement, again resulting in an additional hydraulic head measurement. The black circle (hydraulic conductivity) and crosses (hydraulic head) indicate the previously fixed measurement locations. The white rectangle marks the source location and the white diamond marks the target location (compare Fig. 5.2).

entropy. Therefore, the reverse analysis tends to underestimate data impact for critically small ensemble sizes.

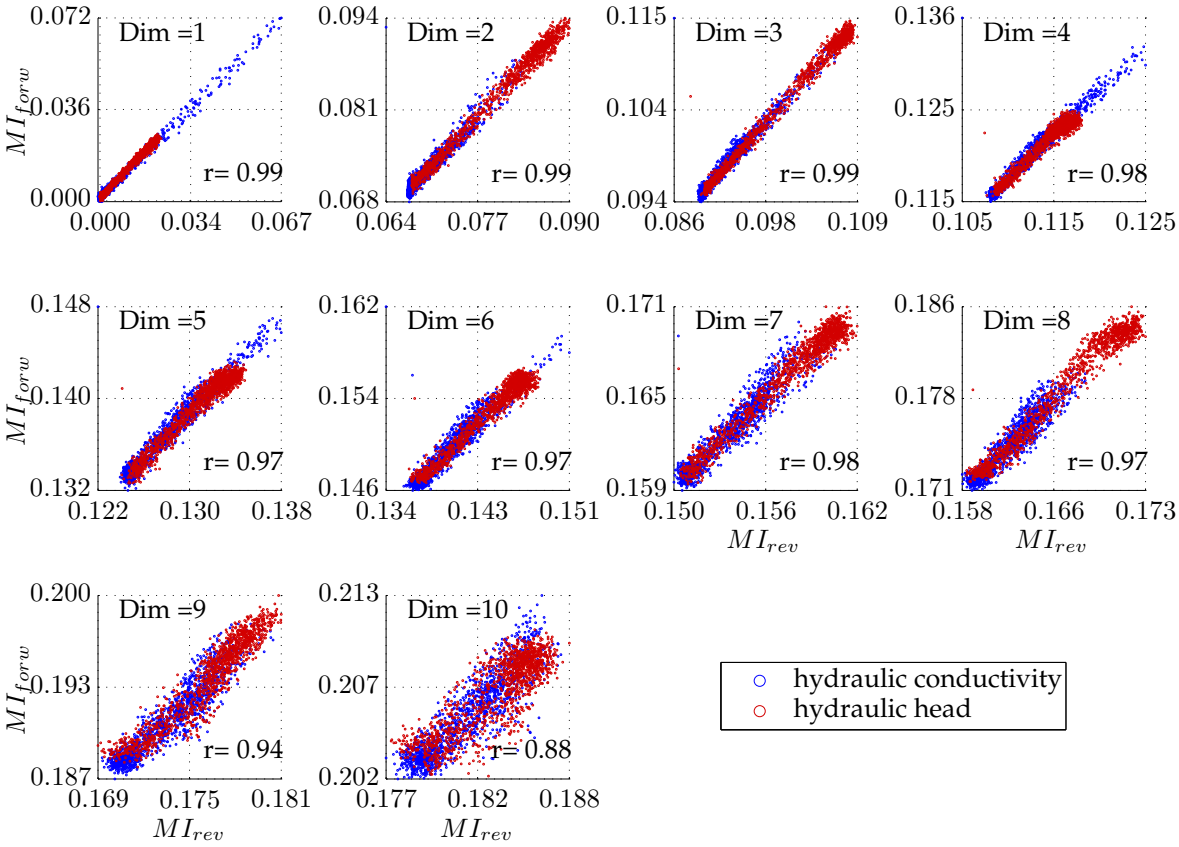


Figure 5.4.: Scatter plot comparing the data impact values obtained from data worth maps in the reverse and in the forward analyses for the design size (dimension) one to ten (TC1). Red points indicate measurements of hydraulic head, blue points indicate measurements of hydraulic conductivity.

Despite the convergence-based bias, the maps show the same characteristics and features: The bias does not significantly change the spatial structure of the data worth maps. Hence, the bias does not change the ranking of designs by the utility function, and the search for the optimum remains unaffected.

To assess whether the ranking in fact remains the same, Fig. 5.4 shows scatter plots comparing forward-reverse pairs of data impact taken from data worth maps at the same location in the domain. Both hydraulic conductivity data (blue points) and hydraulic pressure head data (red points) are shown. The regression coefficient R between both analyses is given for each design size. With growing dimensionality of the design problem, increasing noise between the two implementations can be seen. The reason is the decreasing statistical approximation quality with increasing dimensionality when using a constant ensemble size. The different slopes (visible in the different scales for the X-axis versus the Y-axis) are caused by the bias characteristics

described above.

Overall, the numerical evidence shown here supports the theoretically derived equality also for the two different numerical approximations. In specific, the ranking of data impact is mostly unaffected by the different types of occurring statistical discretization effects. Thus, the main features of the data impact maps are equal for the two analysis directions, which is crucial to reach almost identical near-optimal solutions in both implementations.

TC2: Global optimization

Tab. 5.5 the shows results from a global optimization for a set of ten measurements in TC2 using the genetic algorithm (GA) described in Sec. 3.2.2. The table contains, for both analyses, the highest (best), average (mean) and lowest (worst) re-computed data impact out of the twenty runs. In brackets, I also provide the values expressed as a percentage relative to the highest data impact found altogether.

The optimization of the ten-poin design provided the following results. The forward analysis leads to a best optimal design (out of 20 repetitions) with a data impact of 0.2084. The highest achieved data impact provided by the reverse formulation is 0.2033, which is very close (97.5%) to the forward result. The same proximity can be observed in the averaged values (over 20 optimization runs) of data impact as well. The worst case in the forward formulation provides only 81.9% and is an outlier, for which I assume that it is a non-behaving optimization result of the GA. Omitting the outlier would result in a new worst case of 91.1%. These results show that the uncertainty in a global search algorithm and the differences between both implementations are comparable each other. Also, it shows that the reverse analysis performs well and produces near-optimal solutions just like the forward one.

	Forward	Reverse
best	0.2084 (100%)	0.2033 (97.5%)
mean	0.2024 (97.0%)	0.1951 (93.6%)
worst	0.1708 (81.9%)	0.1894 (90.9 %)

Table 5.5.: Ten-point design data impact expressed as relative MI and as percentage compared to the best value (TC2).

TC3: Evaluation time

This section focuses on the computing times for both implementations. Achieving a substantial speedup was the primary purpose of introducing the reverse formulation. The computation time is dependent on the number of measurements, and I therefore analyze the speedup behavior over design sizes from one to ten (Fig. 5.5). Designs are taken from TC2 to measure the average evaluation time for a data impact assessment (one call of the utility function) with both implementations. I also consider different possible numerical techniques for density estimation, as introduced in Sec. 2.2.3. Below, I will discuss how the observable speed-up is related

to the specific advantages and challenges of the two analyses. The used CPU and GPU architecture and hardware for the speed test is given in Appendix B.

Fig. 5.5 shows the different evaluation times for an ensemble size of 5×10^5 (upper part) as used in TC1 and 1×10^5 (lower part) as used in TC2. The forward analysis is shown as blue lines and the reverse analysis as red lines.

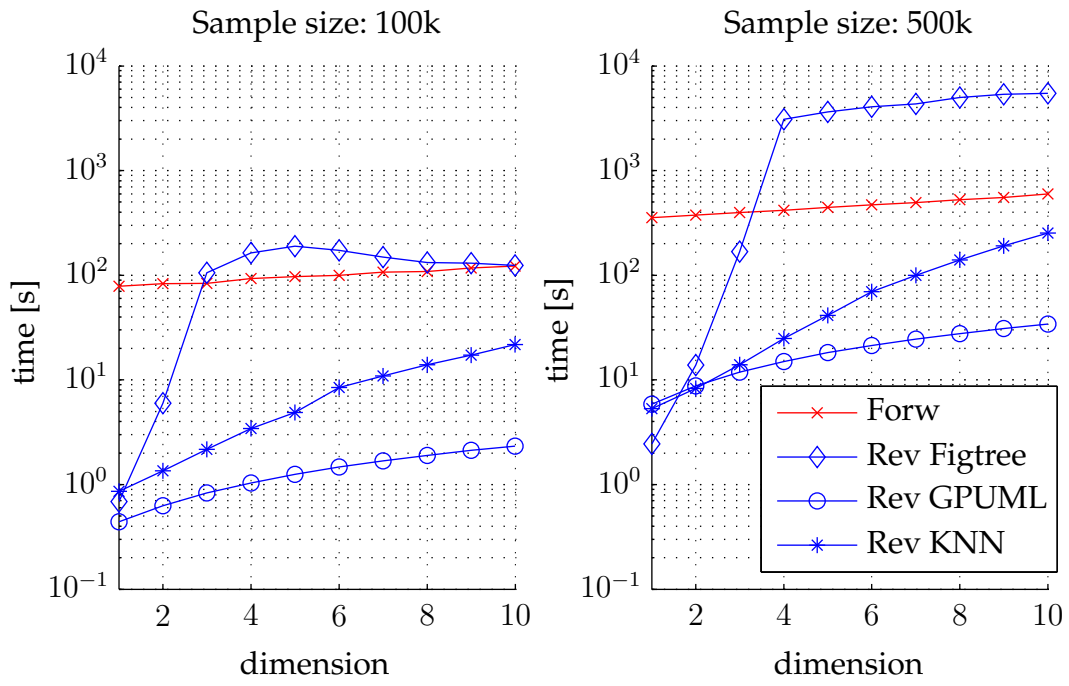


Figure 5.5.: Average evaluation times for the forward implementation (red line) using the fastest implementation and reverse implementations (red lines) of mutual information. The reverse implementations differ by their applied *pdf* estimator. The diamonds (♢) indicate the estimation based on the Figtree library, the circles (○) represent the use of the GPUML library and the start show the times by using the KNN approach. The left figure shows the evaluation times for an ensemble size of 5×10^5, the right part for a size of 1×10^5.

The main computational task in the forward analysis is the high-dimensional conditioning on the measurements, which needs to be repeated over a large number of possible, yet uncollected measurement values. The red line marked with crosses shows the evaluation times when using the CPU-based implementation for KDE-based *pdf* estimation and the weights are evaluated according to (Appendix A(i)).

For the reverse formulation, the blue line marked with the diamonds (♢) shows the evaluation times based on the KDE-based *pdf* estimation. The used implementation is based on the FigTree library by *Morariu et al.* [2008] (also see Appendix A(ii)). The blue line marked with the circles indicates the evaluation times for a GPU-based implementation of KDE using the

GPUML library provided by *Srinivasan and Duraiswami* [2010] (also see Appendix A(iii)). The blue line highlighted by the stars (*) shows the evaluation times when the k-th nearest neighbor (KNN) method is used for estimating the entropy [*Singh et al.*, 2003]. In contrast to the previous ones, this technique directly evaluates entropy for the neighboring distance, without explicitly evaluating the *pdf*.

Among all tested implementations of the reverse implementation, the GPU-based implementation outperforms all other implementations in almost any case. Its use, however, requires a graphical processing unit that is configured to execute mathematical calculations. For this available architecture, the reverse formulation is in any case faster by an order of magnitude and can reach up to two orders of magnitude for small designs.

In case that only a CPU is available, the KNN estimation of entropy still leads to a speedup compared to the fastest available implementation of the forward formulation between one and two orders of magnitude. However, less speedup is achieved compared to the GPU-based evaluation, especially for higher dimensions. The KDE-based implementation does outperform the forward formulation only for small design sizes, but performs best for the data impact estimations of a single measurement. Overall, this shows the advantages of the reverse formulation to improve the evaluation speed for the estimation of data impact.

5.7. Reverse approximated measures of data impact

So far, I showed how the reverse mindset allows evaluating a fully equivalent alternative calculation of Mutual Information. Despite the theoretical potentials of the reverse framework, the involved multivariate entropy estimation for $p(\mathbf{y}|z)$ may pose a challenging numerical problem. This holds especially true when $p(\mathbf{y})$ is high-dimensional, i.e., when many data points are optimized at once and the dimension of $p(\mathbf{y}|z)$ is correspondingly large. However, there are systems and problem classes in which approximated measures may be sufficiently accurate or where time pressure makes their use mandatory. For that reason, this section introduces approximated measures that evolve from this mindset. The key component of such approximated reverse formulations is an efficient approximation of conditional entropy.

5.7.1. Parametric density and entropy estimation

The key idea followed here is to obtain conditional entropies and corresponding MI values without numerical *pdf* estimation. Parametric density estimation is a common approach to avoid the numerical estimation process, but leads to inaccuracies in cases that the real shape of the *pdf* does not match the assumption.

Under the assumption that conditional distribution of the observation quantities $p(\mathbf{y}|z)$ shows a multi-Gaussian dependency, the exact value of the conditional entropy can be calculated based on the covariance matrix $\mathbf{Q}$ of the distribution. This approximated entropy is called Gaussian Entropy h_G in the following and is calculated according to *Cover and Thomas* [2006] in the following way:

$$h_G(\mathcal{N}_{n_d}(\mu, \mathbf{Q})) = \frac{1}{2} \ln(2\pi e)^{n_d} \det \mathbf{Q}, \tag{5.18}$$

where $\det$ denotes the determinant applied on the covariance matrix $\mathbf{Q}$ and n_d is the number of dimensions. For a single measurement, the variance itself is a scalar quantity that can be used as an objective function.

Replacing the (conditional) entropy in the reverse formulation by a Gaussian Entropy approximation leads to a reverse approximation of Mutual Information MI_G. The conditioning is applied exactly as in Sec. 5.4.3. Only the density estimation step is simplified by an analytical estimate based on the covariance. The conditional (co-)variances $\mathbf{Q}_{\mathbf{yy}|z}$ are evaluated for each subsample $z \in Z_n$. The analytical Eq. (5.18) for Gaussian entropy leads to the conditional entropy of each subsample j:

$$\hat{h}_{G,j}(\mathbf{y}(\mathbf{d})|z \in Z_n) = \frac{1}{2}\log((2\pi e)^{n_d}|\mathbf{Q}_{\mathbf{yy}|z} + \mathbf{R}_\mathbf{y}|) \tag{5.19}$$

where $\mathbf{R}_\mathbf{y}$ is the covariance of the measurement errors $\epsilon_\mathbf{y}$. The prior Gaussian entropy is evaluated accordingly. Like in Sec. 5.4.3, the evaluation of the preposterior expectation is averaged over all subsets of Z_q, which leads to the following approximation of data impact:

$$\hat{\phi}_{rev}^{MI_G}(\mathbf{d}) = \hat{h}_G(\mathbf{y}(\mathbf{d})) - \frac{1}{q}\sum_{j=1}^{q}\hat{h}_{G,j}(\mathbf{y}(\mathbf{d})|z \in Z_j) \tag{5.20}$$

Using this parametric entropy estimate simplifies the numerical effort tremendously, as a numerical density estimation is no longer required. Instead, the underlying assumption is that each conditional joint distribution of measurements $p(\mathbf{y}|z)$ is multi-Gaussian and therefore the entropy estimation requires only their covariance $\mathbf{Q}_{\mathbf{yy}|z}$. In addition to this much easier estimation process, the required sample size to estimate a stable covariance is much smaller than the one required for *pdf* estimation.

5.7.2. Numerical test case

In this section, I apply the parametric reverse estimation of MI via MI_G(see Eq. (5.16)) to the test case from Sec. 5.6 and assess the loss of estimation quality compared to the fully nonlinear evaluation. The result is shown in Fig. 5.6 and indicates that the approximation performs well in lower dimensions. The reason is that, in this particular system, the nonlinear dependencies mainly arise within the model relation for the prediction z, while the multi-Gaussian assumption between measurements is adequate. For higher dimensions, increasingly strong errors can be found especially for head measurements.

Fig. 5.7 shows the same analysis for the EnKF-based data impact estimation (as introduced in Sec. 3.3.1) for a direct comparison. The EnKF additionally approximates the relation the conditioning procedure based on data-prediction covariances $\mathbf{Q}_{\mathbf{y}z}$, while reverse approximated MI performs accurate conditioning. The comparison to the full nonlinear Mutual Information shows the inaccuracy of this approximation. Starting from the first dimensions, head measurements are systematically underestimated, as the nonlinearity is partly invisible to linearized approaches. These results are in accordance with the comparison in Fig. 4.8.

The direct comparison of Fig. 5.6 and Fig. 5.7 shows clear evidence, how the partly nonlinear assessment of the reverse approach shows much better approximation quality. Apparently, the

assumption of conditional multi-Gaussian distributed data is less severe than the underlying unconditional Gaussian assumptions for fully-linear data impact estimates (see Sec. 3.3.1).

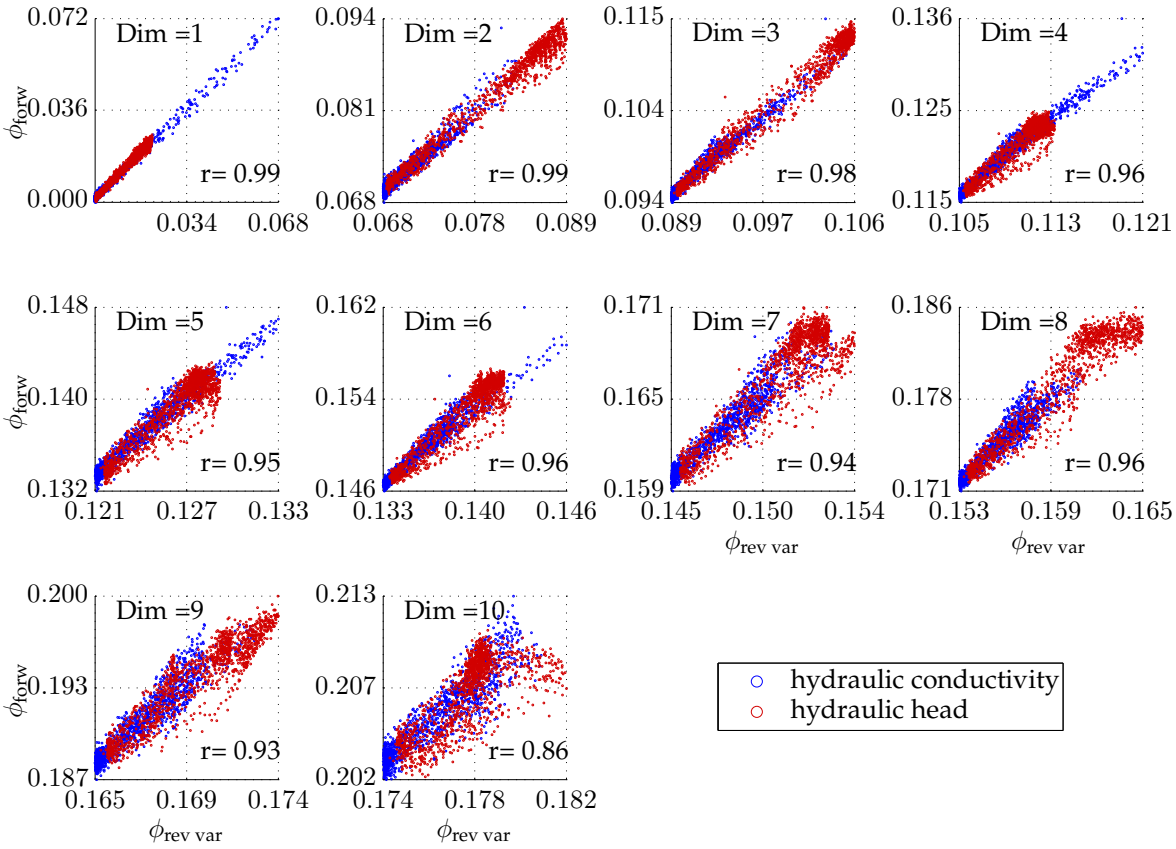


Figure 5.6.: Scatter plot comparing MI (Eq. 5.16) and the reverse covariance-based approximation (Eq. 5.20) for the design size (dimension) one to ten within TC1. Red points: hydraulic head; blue points: hydraulic conductivity.

Finally, the same test case allows to assess the reduced computational time achieved by the approximation. Fig. 5.8 compares the evaluation times of the covariance-based approximation of MI to the times for the reverse evaluation of full MI (compare Fig. 5.5). The evaluation times achieved with the reverse approximated entropy are again one order magnitude smaller than the fastest non-parametric and accurate implementation of reverse MI.

The main reason is, as already mentioned, the absence of the need to estimate a high-dimensional *pdf*. Since the GPU architecture is specifically powerful for the fast evaluation of *pdf*s, the possibility to outsource the covariance approximation to the GPU was not considered.

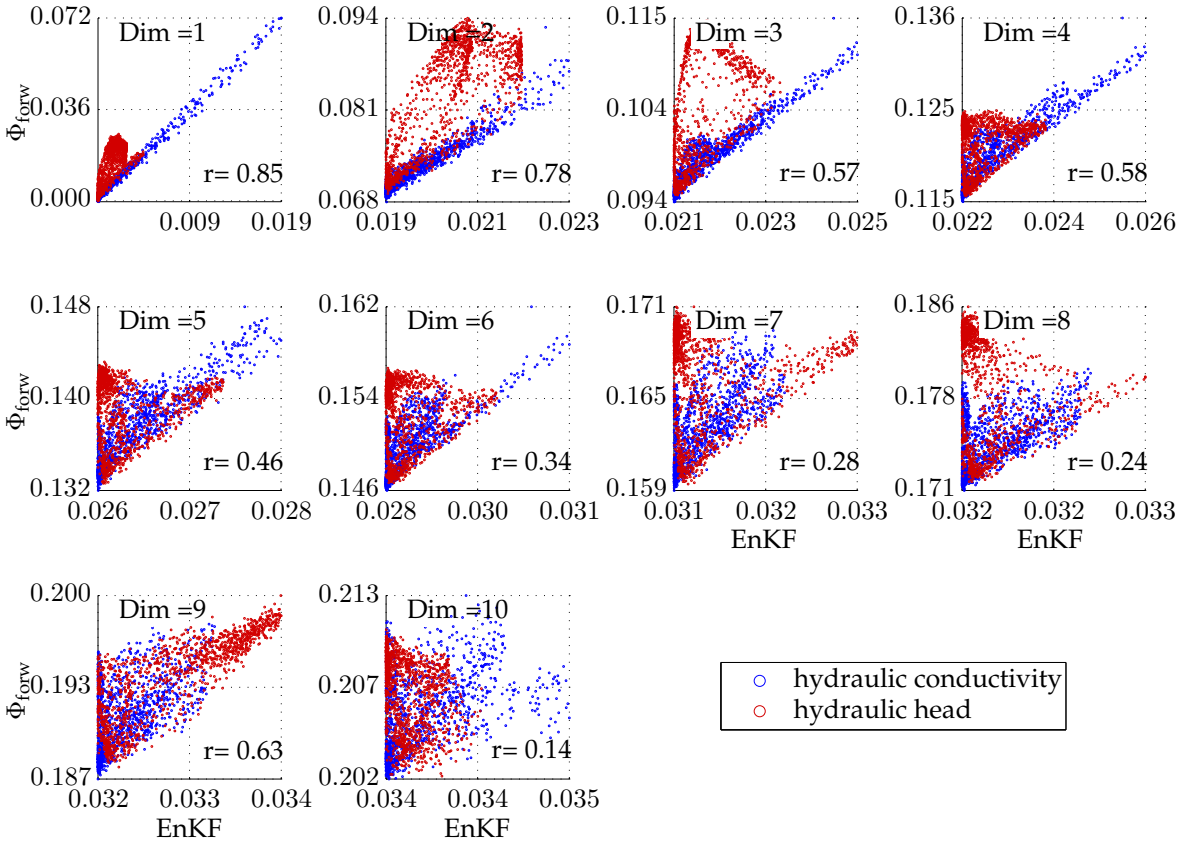


Figure 5.7.: Scatter plot comparing MI (Eq. 5.16) and the EnKF-based approximation (Eq. 3.5) for the design size (dimension) one to ten (TC1). Red points: hydraulic head; blue points: hydraulic conductivity.

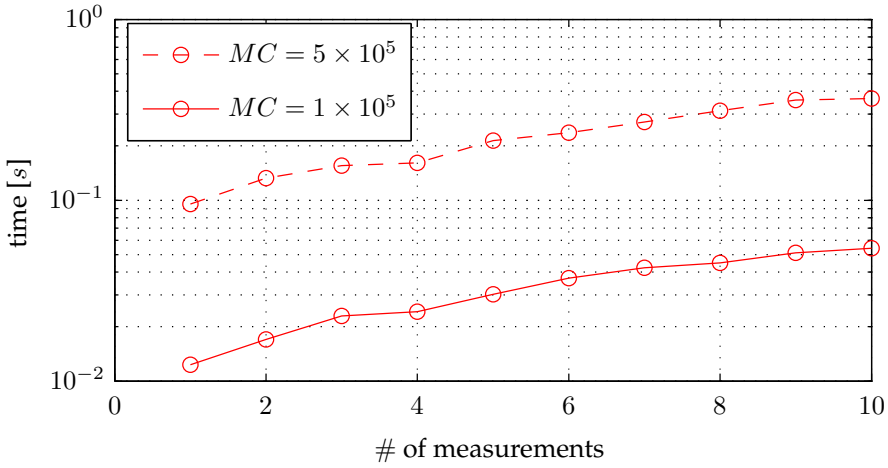


Figure 5.8.: Computation times of the parametric approximation of reverse Mutual Information. Without the need of numerical *pdf* estimation, this approximation is much faster, but potentially inaccurate.

5.8. Summary and conclusion

Summary: In this chapter, I developed a reverse mindset and methods for data impact analysis that can be shown to theoretically yield fully identical optimal designs. In addition, I illustrated this equality in a discrete example. Further, I highlighted the numerical challenges in both conventional and reverse analyses and derived a high speed-up potential for the reverse analysis. In the second part of this chapter, I developed an actual implementation of the reverse analysis and compared it to the forward implementation in terms of quality and speed. In Sec. 5.7, I developed a parametric and thus partly approximated measure for data impact within the reverse mindset. This is based on the analytical evaluation of entropy for multi-Gaussian distributions. The implementation provides significantly better results compared to fully linear measures as the one based on the EnKF at comparable evaluations times. This concludes **Step II** within this thesis.

Overview of data impact estimates The purpose of this section is to give the reader a quick overview of the data impact estimates within this thesis. The methods are listed in Tab. 5.6, starting with the simplest method based on the FOSM approach on the left and listing increasingly more complex methods towards the right. The table lists the used analysis type for statistical dependency, the type of Bayesian update, the applied uncertainty measure and the required assumptions for each approach.

In addition, the individual implementations within this thesis are listed. Except of the FOSM-based approach, all mentioned methods are ensemble-based and therefore rely on converged

ensemble statistics. The list is organized so that the severeness of assumptions is increasing towards the left. This coincides with an obvious reduction in the computational costs.

	Assumptions			Compuational costs	
Data impact estimation	FOSM-based	EnKF-based	Reverse variance-based	Forward Variance-based	Foward / Reverse MI
Evaluation of statistical dependence	local sensitivity	global sensitivity	conditional pdf	conditional pdf	conditional pdf
Bayesian update	linear	linear	nonlinear	nonlinear	nonlinear
Uncertainty measure	second order variance	second order variance	Gaussian entropy	second order variance	entropy
Required assumptions:	multi-Gaussian distribution p($\mathbf{y}$,$\mathbf{s}$,z)	multi-Gaussian distribution p($\mathbf{y}$,z)	conditional multi-Gaussian distribution p($\mathbf{y}$|z)	-	-
Implement-ation within this thesis	-	$\Phi_{\text{EnKF}} = \mathbf{Q}_{z\mathbf{y}}\mathbf{Q}_{\mathbf{yy}}^{-1}\mathbf{Q}_{\mathbf{y}z}$ section: 3.3.1	$\Phi_{\text{Rev var}} = h_G(\mathbf{y}) - h_G(\mathbf{y}\|z)$ section: 5.7.1	$\Phi_{\text{Forw var}} = E_{\mathbf{y}}[V(z) - V(z\|\mathbf{y})]$ section: 4.3	$\Phi_{\text{Forw MI}} = E_{\mathbf{y}}[h(z) - h(z\|\mathbf{y})]$ section: 5.3 $\Phi_{\text{Rev MI}} = E_z[h(\mathbf{y}) - h(\mathbf{y}\|z)]$ section: 5.3

Table 5.6.: Summary of applied data impact estimates within this thesis. The linearization within the estimation raises to the left side, whereas the estimation complexity is higher to the right.

This list of methods for data impact analysis leaves the modeler with the choice which level of complexity is adequate for the currently investigated system. This choice needs to be adequate for the underlying system model to prevent severe approximation errors.

Conclusions: From this combined theoretical and numerical investigation, I conclude the following:

(1) The reverse formulation for nonlinear data impact assessment is theoretically identical to the classical forward analysis. The numerical implementation of the reverse approach provides design quality values that are comparable to those of the forward formulation.

(2) The theoretical speedup potential leads to a practical speedup in most numerical test cases. The fastest implementation showed at least a speed-up of one order of magnitude and

up to two orders of magnitude for small ensemble sizes and low numbers of measurements

(3) The choice of the numerical technique for high-dimensional *pdf*-estimation is essential to optimally exploit the speedup promised by the theoretical considerations. In this context, GPU-based kernel density estimators performed best. The KNN-based entropy estimation performed best among the CPU-based estimators.

(4) In general, the new reverse approach allows modelers to apply nonlinear estimation of data impact within DoE for larger and more complex, realistic problems.

(5) The approximation of the data as multi-Gaussian yields tremendously shorter evaluation times and a much lower required number of realizations. Simultaneously, it showed in general a significantly higher design quality compared to the fully linear approach based on the EnKF. This is the case because the reverse approximation is able to assess parts of the nonlinear dependencies of the system.

6. Interactive design of experiments

This chapter resembles **Step III** within this thesis and approaches a fundamental problem within simulation-based DoE of data acquisition. As mentioned in Sec. 4.2, any optimized decision strategy is based on the current (available) model, which is considered to suffer from unacceptably large uncertainty. Otherwise, one would not desire to acquire more data and to optimize their acquisition. Yet, this uncertain model is the basis of data impact estimation and is employed multiple times within the estimation and optimization of data impact. In fact, one can regard the suggested optimal data acquisition design as yet another uncertain prediction output of the inaccurate model. As long as DoE is considered to be a time-consuming and computationally expensive process and executed before the data collection, this dependence on the current knowledge state is inevitable. However, the fast implementation achieved in **Step II** allows to consider a conceptually new approach to improve the robustness against surprises brought by new data. A sufficiently fast design process would allow interactive integration of the design process within the process of data acquisition and model updating. Conceptual considerations and two actual interactive approaches are introduced in following sections.

6.1. Introduction

DoE of data acquisition for model calibration relies on a model-based utility function, which is evaluated as an output of an uncertain system model. The prior model represents the current belief about the modeled system and is therefore subject to structural, parametric and conceptional uncertainty. The utility function based on this model represents its corresponding information needs, but at the same time suffers from the very uncertainty that poses these information needs. Despite the accurate incorporation of all sources of uncertainty to avoid making wrong claims of knowledge, the model (and hence also the model-based optimal designs) may still be subject to possibly erroneous and wrong assumptions. Hence, the actual measurement and data impact values of the unknown reality will probably be included in the model as a low-probability event, if at all. In other words: Through setting up a model intentionally equipped with uncertainty (and for good reasons), one just admitted explicitly that the model needs improvement and that any model-based quantity (including any found optimal design) needs improvement as well.

Nevertheless, the uncertain model is the currently best prediction model and is therefore used for the estimation of future data impact. The dependence of decisions concerning data collection on an initially unsatisfying model is the key motivation in the current chapter, and will be discussed next (Sec. 6.1.1), followed by a review of the currently sparse state of the art in this field (Sec. 6.1.2).

6.1.1. Prior model dependency

As shown in previous chapters, dealing with complex and nonlinear models requires adequately complex estimates of data impact. This leads to an extensive analysis of the prior model (see Chap. 4) as the most appropriate representation of the current belief system. It introduces a high dependency of the obtained estimates and related design decisions on the current prior. Specifically, there are four aspects to be recognized as to how the prior model is used in nonlinear data impact estimation:

- The prior model is used to encode the general initial state of knowledge and uncertainties in describing the system via prior assumptions and distributions using uncertainty quantification methods.
- Parameter uncertainty is propagated via the prior model to relevant to prediction quantity, which is relevant to define information deficits and needs for task-driven approaches.
- The prior model is currently the best representation of the system, and therefore the most reliable available source of future data. It is therefore used to anticipate the distribution of the (yet uncollected) data values $p(\mathbf{y_d})$.
- For each anticipated potential set of data, the prior model is used in an inverse mode to assess the (nonlinear) impact of proposed designs, based on Bayesian inference.

It needs to be stressed again that , if properly set up, the prior model is the best available representation of the system, given all a-priori available information. For this reason, it is difficult to argue against its usefulness within the estimation of data impact. However, the prior model is subject to several uncertainties that are perceived as being to large; otherwise one would not plan to collect additional data. These uncertainties are described in Sec. 2.3.1) and contain:

- parametric uncertainty,
- structural uncertainty,
- conceptual uncertainty (e.g., initial and boundary conditions or model choice),
- observation uncertainty and
- uncertainty in the simulation of future values of measurement data.

So, in summary, the model-based utility function is a useful (and best available) estimate of data impact, but is subject to the same uncertainty as any other model output. This reveals an inherent vicious cycle for model-based DoE. The less one knows and the more one needs additional information, the less the current knowledge is useful to aid in identifying and collecting the required information.

6.1.2. State of the art

There exist only few approaches that directly address this problematic dependency of DoE on the prior. One general approach for dealing with uncertain utility functions is optimization under uncertainty, which is reviewed by *Sahinidis* [2004]. Examples are *Conditional Monte-Carlo simulation* [e.g., *Maxwell et al.*, 1999; *Copty and Findikakis*, 2000] and *chance constraints*

[e.g., *Tung*, 1986; *Bear and Sun*, 1998; *Freeze and Gorelick*, 1999]. The noisy genetic algorithm [*Gopalakrishnan et al.*, 2003] is an adapted genetic algorithm, which is designed specifically to effectively optimize uncertain utility functions in an expected sense. In addition, multi-objective frameworks, like *robust optimization* [e.g., *Zang et al.*, 2005] introduce robustness against uncertain parameters as a second utility function. Multi-objective search algorithms like, e.g., the BORG algorithm [*Hadka and Reed*, 2013] could be used to find the so-called pareto optima that offer trade-off solutions between highest expected utility and least sensitivity to uncertainty.

All approaches in the field of robust design trade some of the achievable expected data impact for more robust and reliable performance is less favorable cases. This means that expected performance loss is the price to pay for being robust against uncertainties. Furthermore, all approaches are only able to consider robustness against parametric and structural uncertainties encoded in the prior. The methods, however, fail to consider the possibility of changing priors and surprising (low-probability) measurement data, which cannot be modeled easily. Therefore, the approach in this chapter follows a conceptually different approach.

6.2. Approach

The key idea in this chapter is to interactively include the data collection and model updating in the design process within DoE. The approach uses the prior model to optimize data collection, then collects and uses a small fraction of these data , updates (and thus improve) the model , and then re-iterates the optimization for the remainder of data collection in small successive steps. This idea allows for two, fundamentally different, potentials to improve the overall benefits of the data collection and to make it more robust against uncertainty in the the initial prior model:

(a) **Retrospective potential** is the potential of working with a successively improved model in the DoE. The interactive coupling leverages the earliest possible use of data subsets, such that the model and the remaining design process improve progressively with the newly available data. Later stages of the interactive designs become more specific to the updated and hence changing information needs. This reduces the dependency on the prior model assumptions. Overall, the model improves successively, which leads to superior and more specific field acquisition designs.

(b) **Prospective potential** is the potential that is generated by specifically constructing designs by *anticipating* future retrospective potentials. This requires the utility function to include the effects of model updating on later design stages. However, anticipating such effects requires heavily nested optimization loops and inference, which would require enormous computational resources. The arising type of optimization is similar to recursive stochastic decision problems [e.g., *Shapiro and Homem-de Mello*, 1998; *Shapiro et al.*, 2009]. For example, introducing only two interactive sequences would require evaluating a nested second-stage DoE problem for each combination of design and possible data value outcome (preposterior state) in the first sequence. In the literature, only stochastic programing [*Shapiro*, 2008] follows similar principles, however on a drastically simplified level: It assumes that the uncertainty in the utility function reduces over time inde-

pendent of design decisions. Usually, at most two stages [e.g., *Shapiro*, 2008] and only linear(ized) models are considered feasible. This would also affect the initial sequence to potentially increase the learning within the following design process, in contrast to retrospective learning.

In nonlinear DoE problems (for which one single optimization lasts from hours to days), nested optimization procedures seem infeasible. Therefore, I concentrate on retrospective learning approaches, which can be accomplished relatively fast. In the following, I will show two approaches: Sequential DoE in Sec. 6.2.1 and Adaptive DoE in Sec. 6.2.2.

6.2.1. Sequential design of experiments

As the easiest approach, I suggest using sequential design of experiments (SeqDoE) [e.g., *Federov and Hackl*, 1997; *Ford and Titterington*, 1989; *Hu*, 1998; *Feyen and Gorelick*, 2005] as one approach to introduce the most simple approach for interaction between design and acquisition phase. Instead of optimizing all measurements jointly, only a sub-design of a fixed size is optimized in the first sequence. Then, the data according to that sub-design are collected and used to update the model for each following sequence. SeqDoE was mainly introduced to overcome the complex combinatorial problem [e.g., *Haber et al.*, 2008] of optimizing many measurements jointly. In SeqDoE, each subsequence is optimized independently, which tremendously simplifies the design optimization problem. Similar to Greedy search algorithms, the statistical interdependency between measurements of the current sequence to those of future sequences is not considered anymore. This potentially decreases the overall performance, because earlier design sequences do not consider their possible redundancy with later design sequences. Hence, the overall global character of optimization is lost. The loss of globality becomes obvious in the extreme case that each sequence contains only one measurement. In that case, SeqDoE in fact reduces to greedy search, but wit the differences that data are collected and used immediately, one by one.

Fig. 6.1 (left) depicts how the global aspect of search gets lost with increasing number of smaller and smaller subsequences. However, the earlier incorporation of data leads to more robust and more efficient designs through retrospective potentials. By using independent design sequences, it is not possible to maximize the feedback from early data without losing the global character of optimization. Which one of the those two effects (retrospective learning potential or loss of globality) prevails will remain to be seen later in this chapter.

6.2.2. Adaptive design of experiments

Secondly, I developed a second interaction strategy, to which I refer as adaptive design of experiments (AdpDoE). This adapted scheme aims to benefit from the same interactive use of data during the optimization, similar to SeqDoE. However, it will try to consider all statistical interdependencies between measurements of all subsequences. To do so, at first the full design is optimized as in a non-interactive case. The globally optimized full design provides the best estimate of expected data utility, based on the given (prior) state of knowledge and all its

deficits. In contrast to the classical non-interactive design, however, not all data is collected, but only the most informative sub-design is identified from the full design. Similar to SeqDoE, after the subset is identified, the data is collected and the current model is updated. This resulting intermediate posterior model replaces the current prior model and is then used to adapt or re-optimize the remaining part of the full design in the next loop. This interactive data collection and adaptation is continued until all measurements are collected. By doing so, the AdpDoE approach is able to increase the retrospective potential without losing the global character of the optimization, which is depicted on the right side of Fig. 6.1.

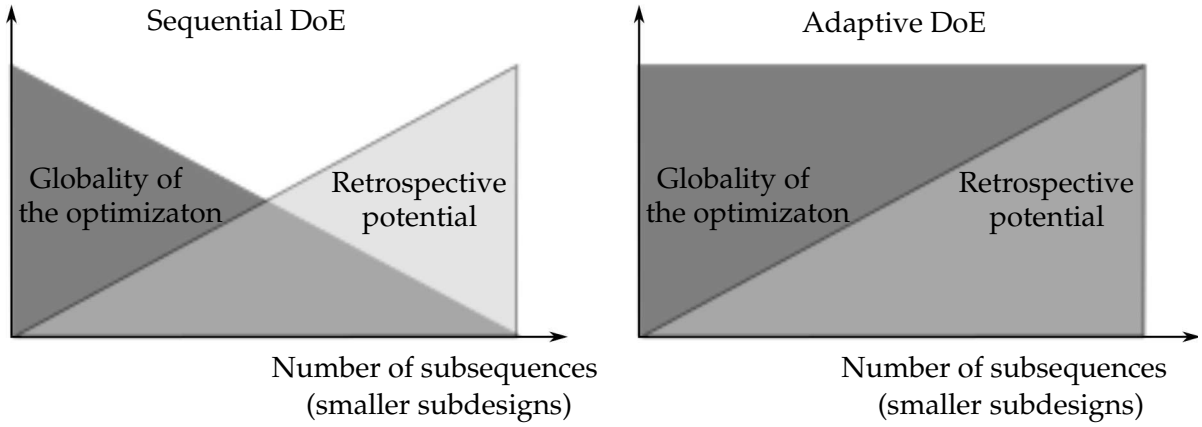


Figure 6.1.: Illustration of the opposing effects of increased retrospective potential through larger feedback frequency for SeqDoE and for AdpDoE.

The mathematical formulations of both adaptive schemes, SeqDoE and AdpDoE, will be provided in detail in Sec. 6.3. The numerical implementations required for the study to investigate the new interactive methodologies are discussed in Sec. 6.4. The synthetic numerical study is introduced in Sec. 6.5 with the goal of comparing the different interaction strategies. I will also address the question of which interaction frequency (number of subsequences) leads to the best overall performance. The results are discussed and compared in Sec. 6.6.

6.3. Mathematic formulation of the interaction methodologies

In this section, I introduce in detail three different strategies, which differ in their interaction level between campaign design and acquisition. All strategies are applied for a DoE problem to optimally design a set of n measurements. A utility function Φ as defined in Sec. 6.4.2 is applied as a common goal of the data collection. This goal (expressed as a function) is also used as a measure of the final success to compare the different methodologies. As discussed, the utility function is dependent on the current model belief $\mathbf{M}_i$, which is now changing during the design process. Therefore, I formally introduce the new functional dependence of the utility function as:

$$\Phi = f(\mathbf{d}_i, \mathbf{M}_i), \tag{6.1}$$

where $\mathbf{M}_i \sim p(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi} | y_0, ..., y_{i-1})$ resembles the current model belief at the design stage i, defined by the distribution of the uncertain parameters conditioned on the already available measurements. In the interactive context, the actual measurement values $\mathbf{y}(\mathbf{d}_i)$ from of the data acquisition executed according to the design stage $\mathbf{d}_i$ is shortly denoted as $\mathbf{y}_i$.

Three different schemes are introduced in the next sections, and are also illustrated in Fig. 6.2.

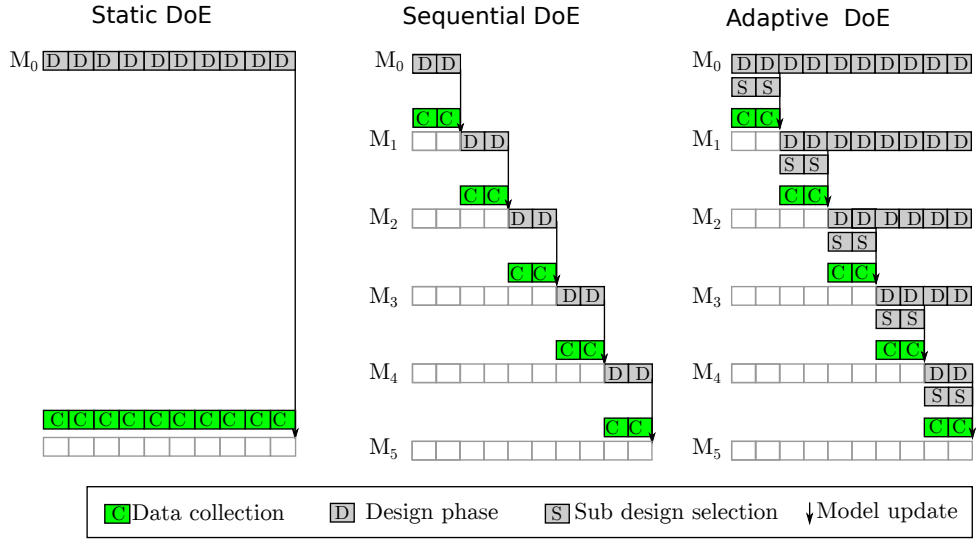


Figure 6.2.: Three interaction methodologies: (1) On the left Static DoE, where the whole design is globally optimized under the prior model $\mathbf{M}_o$ and then data is collected. (2) The center figure depicts Sequential DoE, which uses data from previous sequences for a better design of the following sequences. (3) Adaptive DoE on the right side is also using previously collected data to adapt later sequences, but in contrast to Sequential DoE, in each sequence the full design is optimized jointly to in a global fashion and only a most suitable subset is collected at each stage.

6.3.1. Non-interactive/static design of experiments (StDoE)

The non-interactive DoE (shortly referred to as Static DoE, StDoE) is introduced as the state-of-the-art reference, which does not benefit from any interaction. The design problem in the full design space $\mathbf{D}$ is globally solved once to get the optimal design $\mathbf{d}_{opt}^{St}$ according to:

$$\mathbf{d}_{opt}^{St} = \arg\max_{\mathbf{d} \in \mathbf{D}} \left[\phi(\mathbf{d}, \mathbf{M}_0) \right], \tag{6.2}$$

where $\mathbf{M}_0$ is the prior model and $\mathbf{D}$ is the design space of n measurements. This setup is identical with the general design problem as defined in Sec. 3.1. After that, the data acquisition is executed according to $\mathbf{d}_{opt}^{St}$ and the resulting data values are used for one single model update. In this setup, the optimization and data acquisition is clearly separated and the final quality depends highly on the prior model $\mathbf{M}_0$. The left side of Fig. 6.2 provides the schematic of static DoE for an example with a total of ten measurements.

6.3.2. Sequential design of experiments (SeqDoE)

SeqDoE splits the full $(n_d \times 1)$ design $\mathbf{d}$ in m independent design sequences d_i. Each sub design problem $\mathbf{d}_i$ creates smaller design spaces $\mathbf{D}_i^{Sq}$. In each sequence, the DoE problem is solved independently, by maximizing the design utility.

$$\mathbf{d}_{opt,i}^{Sq} = \arg \max_{\mathbf{d}_i \in \mathbf{D}^{Sq}} \left[\phi(\mathbf{d}, \mathbf{M}_{i-1}) \right] \qquad i = 1, ..., m, \tag{6.3}$$

The optimization is followed by collection of field data according to $\mathbf{d}_{opt,i}^{Sq}$ and the actual measurement values $\mathbf{y}_i$ are used to update the model $\mathbf{M}_{i-1}$ by Bayesian inference of the uncertain parameters $(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi})$. This leads to a new model belief M_i, which is defined as:

$$\mathbf{M}_i \sim p^{\mathbf{M_i}}(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi}) = p(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi} | \mathbf{y}_i, .., \mathbf{y}_i), \tag{6.4}$$

where $\mathbf{y}_k = \mathbf{y}(d_k^{Sq})$ are the data sets collected under the previous design stages $k = 0, .., i-1$. Compared to static optimal design (StDoE), this sequential structure has two major advantages. First, it has the possibility to exploit the retrospective potential by incorporating newly available data, by updating the distribution $p(\mathbf{s})$, $p(z)$ and $p(\mathbf{y_d})$. Second, the sub-designs contain a lower number of measurements, which reduces the dimensionality of the design space in each sequence by a factor of m. Therefore, the search problem as well as the statistical and numerical approximation of the parameter dependencies simplify drastically, leading to much faster optimization. This simplification of the optimization problem was one of the main intentions for the use of SeqDoE in literature [e.g., *Guest and Curtis*, 2009], but simultaneously leads to the loss of the globality of the optimization.

6.3.3. Adaptive design of experiments (AdpDoE)

Adaptive design of experiments (AdpDoE) is set up to benefit from the same retrospective feedback potential as SeqDoE, without suffering from the loss of global optimization. This is accomplished by applying the following steps in each of the m sequences, which are similar to those introduced for SeqDoE. The first step in each sequence i is to optimize the full remaining design problem in a global fashion, which is defined by the design space $\mathbf{D}_i^{Adp}$. This space comprises all remaining measurements. Therefore, initially the full design problem is solved, exactly as in the non-interactive case illustrated in Fig. 6.2. The adaptive design problem for sequence i is generally described as:

$$\mathbf{d}_{i*}^{Adp} = \arg \max_{\mathbf{d} \in \mathbf{D}_i^{Adp}} \left[\phi(\mathbf{d}, \mathbf{M}_{i-1}) \right] \qquad i = 1, ..., m \tag{6.5}$$

The second step in each stage is to select a subset sized n/m of the (remaining) full design. Different selection criteria are possible, depending upon if the modeler wishes to direct or assist the design process by, for instance, focusing on the identification of structural parameters in early sequences. If that is not the case, the intuitive choice is to use the subset that promises

the individually highest (current) expected data impact, measured by $\phi(\mathbf{d}, \mathbf{M}_i)$. The selection of the sub-design in the adaptive sequence of the i-th sequence is therefore formulated as:

$$\mathbf{d}_{opt,i}^{Adp} = \arg \max_{\mathbf{d} \subset \mathbf{d}_{i*}^{Adp}} \left[\phi(\mathbf{d}, \mathbf{M}_{i-1}) \right] \qquad i = 1, ..., m, \tag{6.6}$$

which then is carried out in the field to collect the data accordingly. The data $\mathbf{y}(\mathbf{d}_{opt,i}^{Adp})$ are subsequently incorporated by Bayesian inference, resulting in an updated model $\mathbf{M}_i$ and the respective distributions $p^{\mathbf{M_i}}(\mathbf{s})$, $p^{\mathbf{M_i}}(\mathbf{z})$ and $p^{\mathbf{M_i}}(\mathbf{y_d})$.

The restriction that, in each sequence, the actual design is a subset of the globally optimized design $\mathbf{d}_{i*}^{Apt}$ has the intended purpose to ensure the global character of the optimization. It prohibits the isolation of each design sequence from the remaining optimization, especially for a high number of interaction sequences m.

6.4. Implementation

6.4.1. Ensemble-based Bayesian update

This section describes the implementation of Bayesian updating (as described in Sec. 2.3.5) of the current model after each data acquisition sequence in the interactive strategies. The used updating algorithm needs to be fast and efficient, since real-time capabilities for the interactive coupling with data collection are desired. Similar to the approach in Chap. 4, a single prior ensemble of sufficient size is generated so that it is able to provide a sufficiently large posterior ensemble at each design sequence. Thus, I choose a simple rejection sampling scheme to extract the individual posterior ensemble after each design sequence out of the ensemble that represents the preceding sequences. This avoids the cumbersome generation of new realizations. For the rejection sampling after sampling data $\mathbf{y}_i$ in sequence i, a rejection factor ψ needs to be computed, which is:

$$\psi = \max_{j=1...n_{MC}} \frac{L((\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi})_j | \mathbf{y}_i, .., \mathbf{y}_i)}{L((\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi})_j)}, \tag{6.7}$$

where n_{MC} is the size of the preceding ensemble. This factor represents the maximal likelihood found within the current prior ensemble. The acceptance probability of realization j to become member of the next posterior is computed as:

$$p_{acc,j} = Pr\Big(u_j < \frac{L((\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi})_j | \mathbf{y_i})}{L((\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\xi})_j)\ \psi}\Big)\ ; \quad u_j \sim \mathcal{U}[0, 1] \tag{6.8}$$

The Bayesian update by rejection sampling simply eliminates all realizations according to their likelihood to not comply with the latest collected data. In the result, this is equivalent to working with a likelihood-weighted version of the prior ensemble as in the bootstrap filter. However, rejection sampling reduces the ensemble size and avoids weighted statistics. Therefore, it is much faster than using weighted ensembles based on the BF. This comes, of course, at the cost of slightly reduced accuracy by increased numerical noise. This tradeoff is gladly accepted in the current real-time context.

6.4.2. Utility function

As discussed in the previous chapters, the choice of the utility function to estimate data impact fundamentally influences the computational times to solve the DoE problem. Therefore, interactive DoE strategies can only consider data impact measures that allow fast implementation techniques of data impact estimates. Despite of being the potentially fastest choice, at the time of this thesis, the reverse approaches from Chap. 5 had not been published. Therefore, I chose the forward data impact measure based on variance from Sec. 4.3. The nonlinearly assessed conditional variance has an additional advantage in this context that it requires a smaller posterior sample size compared to the fully accurate evaluation of MI. The implementation of the utility function used in the case study is given in Eq. (4.7). Other choices of utility functions and their impact on the interactive performance will be investigated in the following chapter Chap. 7.

6.5. Application and test

To compare the different DoE/DA interaction strategies, I choose a contaminant transport scenario in a heterogeneous aquifer, similar to the one featured in Chaps. 4-5. In addition to measurements of hydraulic conductivity and hydraulic head, this scenario also considers a measurements of a steady state pumping test. The physical properties of the scenario are described in detail in Sec. 6.5.1. The different test cases, which are considered to evaluate and compare the performance of the different DoE/DA strategies, are introduced in Sec. 6.5.2.

6.5.1. Scenario setup

The featured system is a depth-averaged heterogeneous aquifer with a recent (but from then on continuous) contaminant spill at a known location. The system geometry is sketched in Fig. 6.3, including the location of the source$(x_s, y_s) = (80m, 80m)$, a prediction target and the location of the boundary conditions. The uncertain structural parameters of the uncertain transmissitivity field $\mathbf{T}(\mathbf{x})$ of the heterogeneous aquifer are given in Tab. 6.1. Furthermore, the general flow direction is assumed to deviate from the centerline of the system by an unknown angle ν. Flow boundary conditions are implemented as Dirichlet boundaries to match the randomly drawn values of ν per realization. The transport boundary conditions are set to zero for the domain boundaries and to the maximum concentration c_0 in the source area. All known parameter values and prior distributions for unknown parameters (geostatistical and flow transport parameter) are given in Tab. 6.1.

The prior model setup as described above is used to generate a prior model ensemble that contains structural and parametric uncertainty. This prior model is used to generate 50 random synthetic realities for testing purposes and a prior ensemble of five million realizations for the optimal design of data acquisition (see Sec. 2.3.1).

I consider three different measurements types: hydraulic conductivity T, hydraulic head h and drawdown data d from the pumping test. The drawdown data are from a steady-state pumping

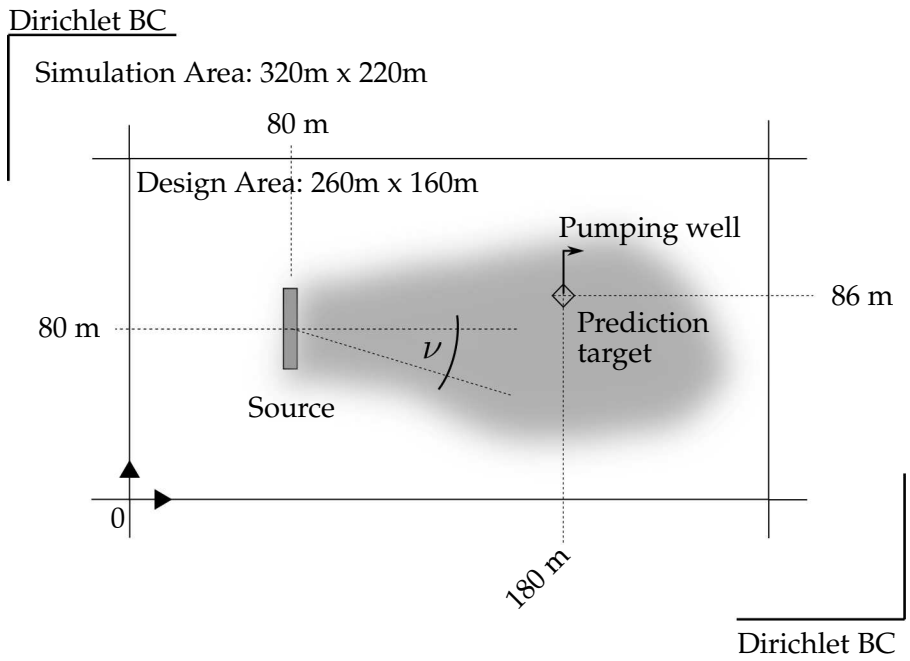


Figure 6.3.: Numerical example, including the location of the boundary conditions, contaminant source location and the target location for the concentration prediction. The design area contains all possible measurement locations that are discretized on a regular grid with two meter spacing.

Transport parameters			
Head gradient	γ	[-]	0.01
Effective porosity	n_e	[-]	0.35
Local-scale dispersivities	[α_l, α_t]	[m]	[0.5, 0.125]
Diffusion coefficient	D_m	[m^2/s]	10^{-9}
Contamination width	l_s	[m]	20

Geostatistical model parameters $\boldsymbol{\theta}$			
Global mean	$\beta_1 = \ln T$	[-]	$\ln 10^{-5}$
Trend in mean	β_2	[-]	0
*Variance	σ_T^2	[-]	$\mathcal{N}(\mu = 2.0, \sigma = 0.3)$
*Integral scale	λ	[m]	$\mathcal{N}(\mu = 15, \sigma = 2.0)$
*Matérn Kappa	κ	[-]	$\mathcal{U}(a = 5, b = 36)$

Boundary parameters			
*Deviation from center	ν	[°]	$\mathcal{N}(\mu = 0.0, \sigma = 10)$

Measurement error standard deviations			
Hydraulic conductivity	$\sigma_{r,T}$	[-]	1.00
Hydraulic head	$\sigma_{r,h}$	[m]	0.01
Drawdown	$\sigma_{r,d}$	[m]	0.02

Table 6.1.: Known and unknown model parameters for the flow, transport and geostatistical mode,l and assumed measurement errors

test with the pumping location being at $(x_p, y_p) = (160m, 86m)$. The considered measurement errors for each type are given at the bottom of Tab. 6.1.

6.5.2. Test case setup

Each strategy is applied to optimally design a total number of ten measurements. Each DoE problem is solved using the genetic algorithm introduced in Sec. 3.2.2. Three test cases with some variations each are performed and summarized in Tab. 6.2:

Test case A1 to A2 is the *non-interactive* reference without any data feedback, according to Sec. 6.3.1. Therefore, the result of TC-A is only one single optimal design, which is based on the prior model. For testing and comparison, this design is finally applied for data acquisition in each of the 50 synthetic realities and the corresponding 50 actual data impact values are calculated. Both a local GS and a global GA are applied as search algorithms.

Test cases B1 to B3 employ the *sequential* DoE/DA strategy accordingly to Sec. 6.3.2. The three variants consider two, five and ten sub-sequences, to investigate the influence of interac-

tion frequency (retrospective potential) versus the loss in globality of the optimization. After the initial design, each synthetic reality provides different measurement values $y_R(\mathbf{d}_i)$, which lead to different optimal designs in later sequences. This will provide 50 different optimal designs, for which the actual data impact is calculated after the final updating step. The use of the GS algorithm in TC-B3, does not effect the globality of the design, since the GS performs a global search for a single measurement.

Test cases C1 to C3 apply the *adaptive* DoE/DA strategy according to Sec. 6.3.3. Again, different numbers of sub-sequences are considered. Apart from the different DoE/DA strategies, the test cases are evaluated in the same fashion as TC-B1-B3.

6.6. Results

In this section, I will discuss the results from the application example. After revisiting the base-case TC-A of non-interactive design in Sec. 6.6.1, I will first point out in Sec. 6.6.2 how feedback from data (i.e., the retrospective potential) is influencing the consecutive design process. Sec. 6.6.3 focuses on the differences between the adaptive and the sequential scheme that arise from the reduced optimization complexity within the sequential approach. It will become clear that dropping measurement interdependencies between measurements leads to a focus on individually strong measurements. In Sec. 6.6.4, I will illustrate the overall performance of the different schemes with respect to the optimization task, using the average performance over all 50 synthetic realities as metric for comparison.

6.6.1. Base case: non-interactive design (TC-A1 - A2)

Before focusing on the feedback effects, it is required to briefly introduce the results of the non-interactive design (StDoE) applied in test case TC-A. StDoE only provides one set of optimal

Reference	(1) Interaction type type	(2) Number of sequences	Search algorithm
TC-A1	StDoE	1	GA
TC-A2	StDoE	1	GS
TC-B1	SeqDoE	2	GA
TC-B2	SeqDoE	5	GA
TC-B3	SeqDoE	10	GS
TC-C1	AdpDoE	2	GA
TC-C2	AdpDoE	5	GA
TC-C3	AdpDoE	10	GA

Table 6.2.: Definition of test cases

measurement locations and types, since no feedback mechanisms apply. The resulting optimal placement of ten measurements is depicted in Fig. 6.4. Five measurements of hydraulic conductivity at the source region are selected for the quantification the contaminant discharge from the source, whereas five hydraulic head measurements close to the boundaries provide information about the ambient flow direction. Overall, these characteristics are similar to those discussed before in Chap. 4, as the test cases resemble each other and no drawdown data were identified to be useful in this design approach.

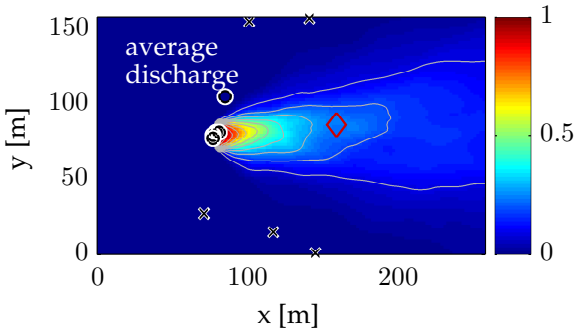


Figure 6.4.: Non-interactive optimal design for placing ten measurements of either hydraulic conductivity (o) or hydraulic head (x), resulting from StDoE. In addition, the average contaminant plume is depicted in the background.

6.6.2. Data feedback effects (TC-B2)

To illustrate the effect of interactive feedback, I chose test case TC-B2, which optimizes five consecutive sets of two measurements each. Of course, feedback effects of data are individually different for each of the 50 synthetic realities. The reason is that, in each of the 50 realities, the later design sequences depend on the actual data valued collected from the specified synthetic reality in the previous design sequences. Thus, the feedback effects cannot be illustrated completely. Instead, they are illustrated by two main characteristics:

- The influence that early knowledge on the source discharge has on the remaining optimal designs of all later sequences.
- The influence that knowledge about the ambient flow direction has on the remaining design sequences.

These aspects were chosen, because the static design in Fig. 6.4 revealed that the source discharge and the ambient flow direction offer a high information value for this given problem.

Early source discharge information: Since the source is defined as a Dirichlet condition ($c = c_0$), the contaminant mass discharge from the source is proportional to the amount of water flowing through the source area ($\dot{m} \sim q_{source}$). Fig. 6.5 illustrates the way how SeqDoE is

adapting subsequent measurement designs, based on the corresponding early information. As an illustration, the synthetic realities are split into realities with the 20 highest and 20 lowest source discharge values. The remaining 10 cases are dropped to intensify the difference between high and low discharge. The left side within the upper row of Fig. 6.5 shows the average contaminant plume for the twenty realities with the lowest discharge, and the right side for the twenty realities with the highest discharge. One can easily see that the higher discharge leads to a wider plume and to a higher chance of the plume to actually hit the target location. The contour lines outline different concentration levels within the contaminant plume. The figure also shows the initial set of two measurements that are located at the source. These two measurements are the result of the optimization in the first sub-sequence, based on the prior sample, still without any feedback. Therefore, this first design sequence is identical for all underlying synthetic realities. The correspondingly collected data values on source-area conductivity, however, differ in the individual synthetic realities. From this point on, the design process receives different feedback from the individual realities and hence evolves differently.

The three lower plots on the left side in Fig. 6.5 show all optimal measurement locations for the low-discharge case in all 20 realities. The resulting measurements from all subsequences are plotted together in one single plot to reveal the common design characteristics that result from the early information on low discharge. The three measurement types (conductivity, heads and drawdown) are separated in individual plots, supplemented by the measurement placement density. The three plots on the right side show the respective information for the high-discharge cases.

For measurements of hydraulic conductivity, the plots in the second row show more intensive sampling of the source region for low discharge. There are two possible reasons for this: First, additional sampling can help to further identify the real source discharge. This information can significantly reduce the prediction variance at the target point in case the source discharge is close to zero. In this case, the prediction variance becomes zero as well. Secondly, the exact position of the small contaminant plumes that emerge from low-discharge source areas is difficult to track far downstream. Therefore, the limited number of allowed measurements can not serve to sufficiently locate the path of the plume. This leads to reduced sampling of hydraulic head measurements (typically used to infer the transverse gradient in TCA, and thus serving to track the plume) as depicted in the third row. Instead, drawdown measurements are preferred in the region around the well. These data provide information about the meso-scale conductivity around the well and therefore about the local flow field around the target location (which is identical to the pumping test well). This helps to infer whether there are local flow focusing effects at the target location, and thus convey the information whether the plume (although relatively thin) might be detoured into or away from the target location.

For high source discharge, and therefore broader contaminant plumes, the respective optimal designs follow different information needs (see right column of Fig. 6.5). The source region is still sampled in a similar fashion by conductivity measurements, however with more focus on the flanking regions. These measurements put the measurements of source conductivity in relation to the neighboring values and help to identify whether the flow focusing effects that lead to high-discharge areas possibly favors the flow from north to south of the centerline. Non-symmetric flow focusing leads to broad plumes that, on average, may have their plume

fringes (rather than the plume center line) aligned with the center line of the domain. The most prominent change of sampling can be identified for hydraulic head measurements, which are almost doubled in their number. Not only the uncertain boundary condition is sampled, but also locations downstream of the source are of importance. The reason is that a broad plume will most likely hit the well location, leading to possibly high concentration values. Therefore, even an approximate path of the plume (as provided by head measurement) offers high information with respect to the target concentration. This helps to distinguish cases in which either the plume center, or only the outer flanks, hit the well location. The local flow characteristics within the well region are, however, less informative for a broad and more homogeneous plume. Therefore, drawdown data are only sampled in very few cases.

Early ambient flow information: The second important information which I consider is the ambient flow direction. As already mentioned, the first sub-design samples only hydraulic conductivity measures. Therefore, the ambient flow direction is only influencing the adaptive designs after the second sequence when hydraulic heads near the lateral boundaries are measured. Again, for the sake of discussion, I distinguish two cases. The first case combines the 20 synthetic realities in which the ambient flow deviates southwards, and therefore away from the prediction target location. This is depicted in the first row of Fig. 6.6 on the left side. It also shows all measurements placed during the first two sequences. The following sequences are influenced by the collected hydraulic head information. In the second case, I separate the 20 realizations in which the ambient flow is deviated towards the target well (see right column of Fig. 6.6. The intermediate 10 synthetic realities are omitted, once again, for the sake of better contrasts and better discussion.

For the sampling of hydraulic conductivity in the source region in the southward-deviating case (see second row, left side of Fig. 6.6), one can identify an increased sampling at the upper flanking locations. The contour lines in the upper left plot indicate the average contaminant plume location. These show that the plume center line is located south of the well. Therefore, the northern flank of the plume is critical to predict the well concentration. This is the reason for the increased sampling north of the source, because the conductivity ratio between these measurements and the source area provide information about the average position of the northern flank. Hydraulic head measurements are further used to estimate the ambient and local flow, see Fig. 6.6 third row, left plot. One might recognize a slight alignment perpendicular to the ambient flow direction. The deviation away from the well also affects the information value of the drawdown data. As the plume most likely passes the well to the south, the local flow field around the well is of low importance.

For a diversion of the ambient flow northwards and therefore towards the well (see right column of Fig. 6.6), the centerline of the plume now bypasses the well in the north. Therefore, the southern flank mainly affects the concentration prediction, and one can observe a correspondingly increased sampling of the lower source region with conductivity measurements. For the diversion towards the well, the alignment of hydraulic head measurements perpendicular to the ambient flow becomes more apparent (see Fig. 6.6 third row right plot), whereas the sampling intensity is similar. Drawdown data is now more important, as the center plume line is deviated towards the well, which is therefore more likely hit by the plume.

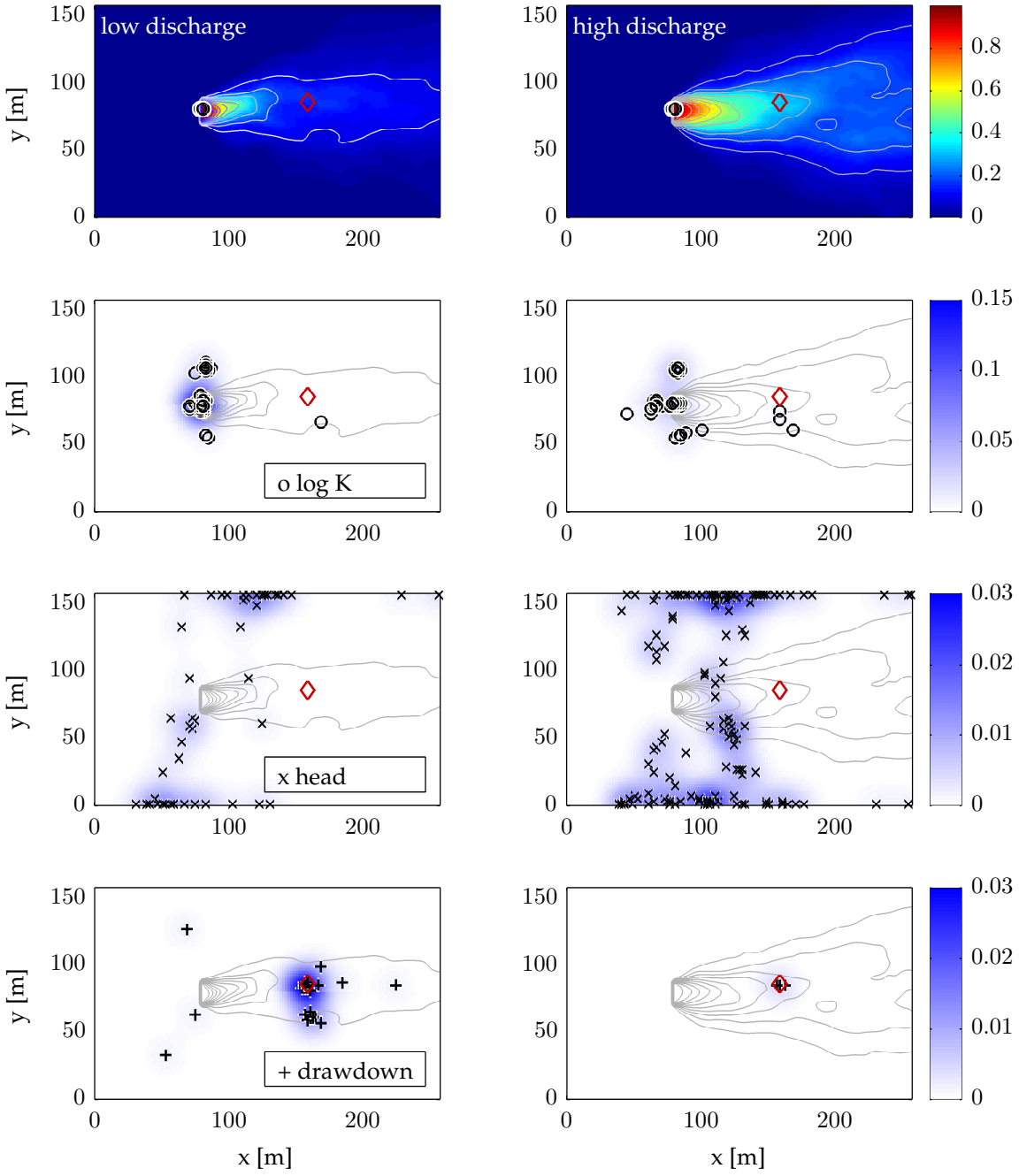


Figure 6.5.: Different optimal design locations for realities of low discharge on the left side and high discharge on the right side. The upper row shows the mean concentration plume for both cases. The lower three rows show the three different measurement types separately.

These two prominent examples clearly show how the feedback from early data positively influences later design sequences. For the other test cases, similar effects can be observed. Of course, the interaction frequency influences the effect of feedback and at which point in time it affects the design.

6.6.3. Complexity of measurement interactions: greedy versus global search

This section investigates the possible disadvantages of SeqDoE versus AdpDoE due to the greedy (rather than global) search algorithm used in SeqDoE. In specific, I analyze the results from the current test case to see whether SeqDoE in fact honors obvious direct data impact more than complexly coordinated sampling patterns.

At first, I consider the number of hydraulic head and drawdown data as an indicator for complex patterns. The idea behind this consideration is that hydraulic conductivity is a more direct type of information. Also, head data are mostly useful to infer hydraulic gradients, and at least two such measurements are required to obtain a gradient. Therefore, greedy search, which places one (few) measurement at a time, can not assess this joint information adequately. A second indicator is the repeated sampling of individual measurements. Clearly repeated sampling is beneficial as multiple measurements of the same quantity reduce the impact of measurement errors in the inference. Still, repeated sampling may also occur if the redundancy of early sampling locations with later sampling locations is not foreseen in the greedy-type search approaches.

Fig. 6.7(a) shows the relative number of the different measurement types. Again, this analysis is averaged over all 50 optimization runs with their synthetic realities. The two different schemes are compared for the case of the ten subsequences (TC-B3 and TC-C3), for which the differences are expected to be largest. The first apparent difference appears in the second sequence, where the sequential approach suggests, on average, significantly fewer head measurements. Similarly, the sampling of drawdown data starts later in the sequential approach. Generally, in later sequences of the sequential approach, the indirect measurements are sampled less than in the adaptive approach. This indicates how the adaptive approach is able to devise more complex strategies.

Fig. 6.7(b) shows the number of repeatedly sampled measurement locations. The figure indicates the higher tendency of the sequential approach to use repeated sampling, which can be interpreted to indicate lower complexity in the individual designs.

In summary, both indicators show the expected decrease in design complexity for the sequential interaction approach. Of course, these effects are less prominent for smaller numbers of sequences, as more measures are considered jointly within the individual sub-sequences. The degree to which this deficit influences the quality of the optimization process depends on the type and characteristics of the underlying model and will be evaluated in the next section.

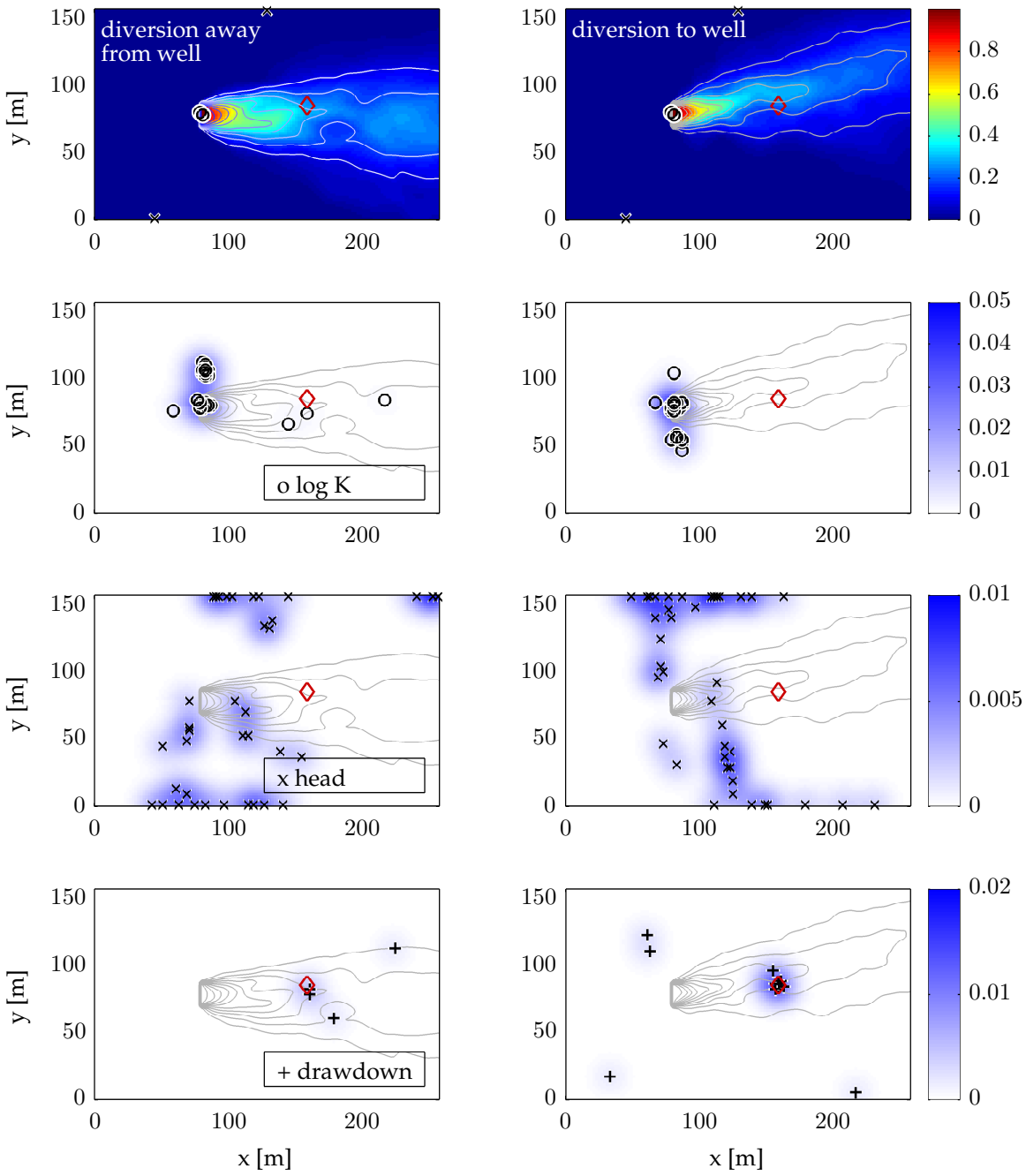
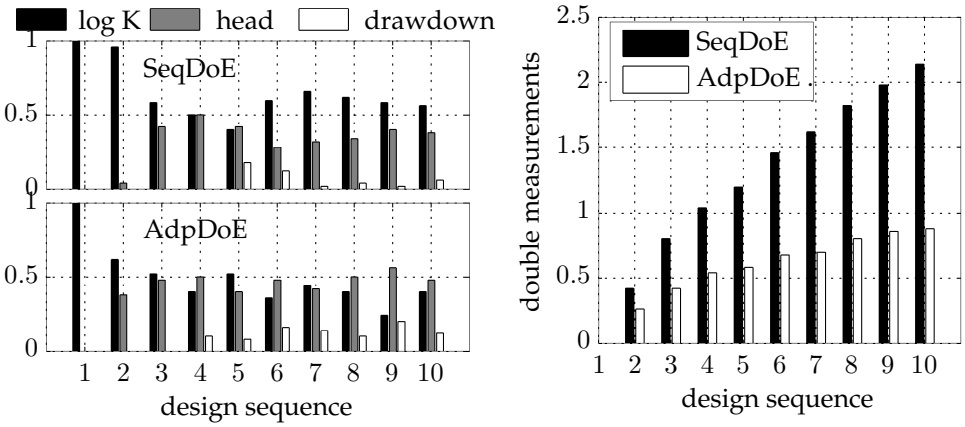


Figure 6.6.: Different optimal design locations for realities in which the plume deviated away from the well on the left side and in which the plume deviates to the well on the right side. The upper row shows the mean concentration plume for both case. The lower three rows show the three different measurement types.



(a) Relative number of samples measurements types compared for the sequential and adaptive strategy for ten subsequences

(b) Number of double placement of measurements compared for the sequential and adaptive strategy

Figure 6.7.: Complexity analysis on the basis of two indicators

6.6.4. Overall performance

Having analyzed of the ability of interactive approaches to exploit early knowledge for more specific optimization in later design sequences (Sec. 6.6.2) and having analyzed the different design complexities between the two interactive schemes (Sec. 6.6.3), the current section discusses their overall performance in all test cases. The applied utility function in all tested approaches is the expected variance reduction in the model prediction. Therefore, the overall performance of a given design after all data are collected (interactive or not) is the final posterior variance. In order to make the analysis independent of the used specific realities, again the average over all 50 synthetic realities is used for comparison.

Fig. 6.8 shows the performance of the different DoE/DA strategies for two (left figure), five (middle figure) and ten (right figure) subsequences. The horizontal dotted black line shows the performance of the best non-interactive design for comparison. On average, the global non-interactive approach using global search reduces the variance by 18%, indicated by the black cross. For completeness, the same non-interactive approach using greedy search is depicted by the black circle. Greedy search only achieves a variance reduction of 17%, illustrating the slight edge of the global optimization on greedy optimization in the given application.

In the left side of Fig. 6.8, the red line shows the posterior variance, averaged over the 50 synthetic realities, of the SeqDoE using two subsequences. Clearly, the interactive approach leads to a major improvement in the overall performance. The reason is that the second design sequence can benefit from the knowledge gained by collecting data according to the first design sequence. The blue line shows the result of AdpDoE, which performs slightly superior to SeqDoE. The more expensive feedback approach reduces the posterior variance by 23%.

The middle of Fig. 6.8 shows the posterior variance development for five sub-sequences. One can observe a slightly different ranking between the three approaches, but an overall improvement of the interactive approaches compared to the previous plot. Apparently, increasing the

interaction frequency is indeed helpful. On average, one can see that already after only three sequences and the collection of six measurements, both feedback approaches show superior or equal performance compared to the non-interactive approaches. This means that the interactive approaches can reach the same overall performance with only 60% of the sampling costs.

The right side of Fig. 6.8 shows the result for applying ten feedback sequences. The sequential approach performs best from the beginning and in the end achieved the biggest variance reduction of 27%. Alternatively, it would allow to provide the same performance than the non-interactive design with only half of the measurements.

Though leading to the more complex designs, the more complex adaptive approach (AdpDoE) does not perform better than the sequential one. In the cases of two and five interaction sequences it performs more or less equally good, but for ten interaction sequences it performs notably worse. In the following, I will discuss this surprising development.

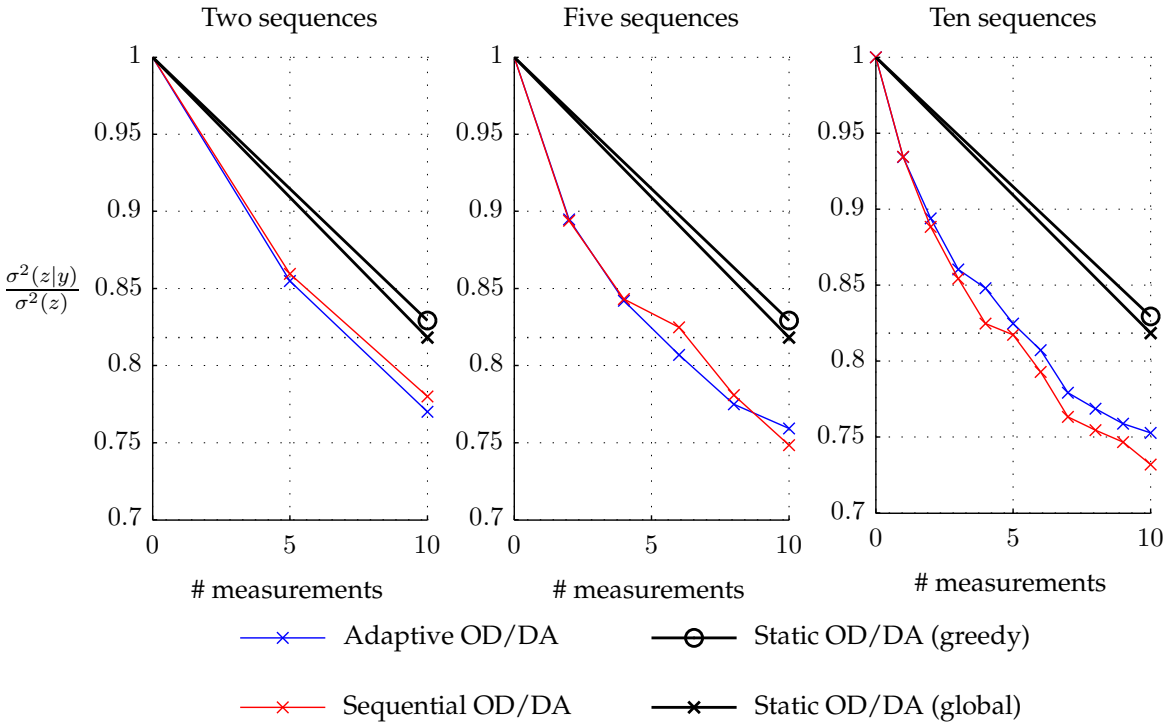


Figure 6.8.: Results of all seven DoE/DA strategies. The blue lines depict all AdpDoE strategies, the red lines show the results of the three SeqDoE strategies and the continuous black line shows the non-interactive StDoE.

6.7. Discussion

The discussion about the surprising performance behavior of the AdpDoE approach can be guided by the question of global versus greedy optimization. Without feedback, the sequential

approach is equivalent to a greedy search algorithm (Sec. 3.2.1), whereas the adaptive approach resembles a global optimization using the GA (Sec. 3.2.2). Both non-interactive strategies are illustrated in the left column of Fig. 6.9, where only the initial model M_0 is the basis for the entire design process. The black line represents the decision path, which is taken during optimization. Since no data become available during the design phase in non-interactive approaches, no change in the design process is illustrated. In that case, it is well known that a greedy-type optimization does not lead to a global optimum and therefore leads to inferior designs compared to a global approach. The superior performance of the global search is also shown by the performance of both approached shown in Fig. 6.8, indicated by the black lines.

In contrast, if data feedback influences the design paths due to the changing model, the later decisions depend on the sampled data values from earlier sequences. This dependency cannot be foreseen by any of the introduced interaction schemes. The initial design sequence is based on the prior model, which is depicted in red for SeqDoE in the upper right side of Fig. 6.9 and for AdpDoE in the lower right side of Fig. 6.9. For this first step, it is clear that the sequential approach provides an equal or superior design quality compared to the adaptive global strategy. The reason for this is that both optimize the same first design, however, AdpDoE optimizes jointly the global design. Therefore, for the initial step, AdpDoE optimizes under the additional restriction that the first design is a subset of the global optimum of all sequences. This additional restriction naturally leads to a data impact which is either equal to or lower than the data impact of the greedy-like optimization in the first step. Only in later stages, the global character of AdpDoE may pay off, due to the better interaction between measurements.

From there on, a race between opposing factors evolves: First, the global character of AdpDoE can possible produce more foresighted and more complex designs that manifest the theoretical advantages of global search. This has been discussed in Sec. 6.2.2. However, the global foresight is based on a model that is outdated one sequence later, after data have been collected. In this aspect, the restriction for the next chosen sampling locations to be a subset of the allegedly (soon-to-be outdated) global optimum can be rather disadvantageous. The right column of Fig. 6.9 illustrates the numerous models states that possibly evolve from different possible future data values. The sequential approach is optimizing the design exactly up to the point at which newly collected data is changing the model belief, which is a robust consideration, not trying to project future decision beyond the validity of the model. The adaptive approach, as introduced here, considers all sequences, but based on the current model. Thus, it ignores the effects of model updates. This is depicted in the bottom of the left column of Fig. 6.9, where the initial design considers one potential path, based on the current model. Therefore, considered interactions between measurements of later sequences, which would make the global design superior, potentially do not pay off.

Therefore, a truly global optimization scheme would require looping over all possible event paths of model states in a recursive manner. Such a scheme would require multiple layers of nested optimizations for only one evaluation of data impact on the first sequence. For the given computational resources, such approaches are far beyond the current possibilities. Even more, these potential paths are only estimated based on the available prior model, so there is no guarantee that such a global optimization would provide improved results.

In summary, the results of this numerical study indicate that, within the interactive DoE strate-

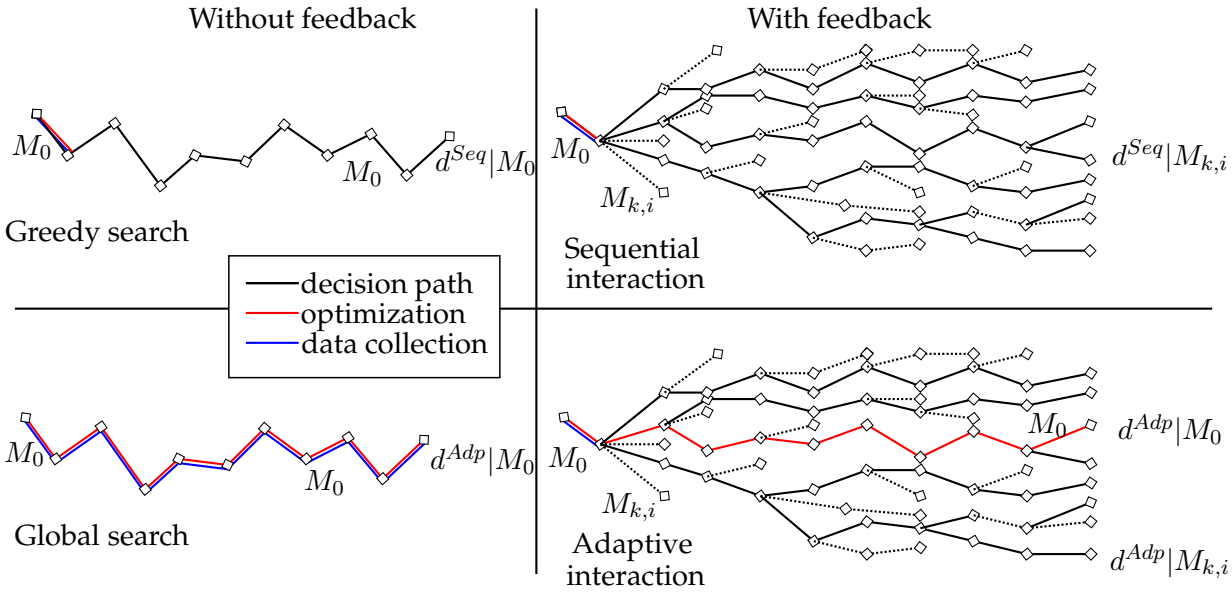


Figure 6.9.: Illustration of the global optimization within the feedback methodology that illustrates how many measurements are considers for the difference cases. Two case are illustrated, assuming the ignorance feedback (no model update) and feedback from data. In case of ignoring feedback, the entire design process is based initial model M_0. For the design process including feedback, the underlying model changes after each Bayesian update.

gies suggested and tested here, the sequential design approach performs best.

6.8. Summary and conclusions

Summary: Model uncertainty is an inevitable factor in the simulation-based DoE of data acquisition. If one desires to have a model flexible enough to adapt to any foreseeable data, the prior needs to be chosen very general to avoid unjustified assumptions. High uncertainty in the available system model directly translates into high uncertainties in the assessment of the data impact, and hence also in the question what is an optimal design to provide the most valuable information.

In this chapter, I introduced two interactive schemes that couple the design of field campaigns and their corresponding data collection. Both schemes were compared in a numerical study within the context of subsurface data acquisition for improving a flow and transport model. In addition to different interaction schemes, different interaction frequencies (or number of measurements per design stage) have been investigated.

On a conceptual level, I showed the challenges and potentials of interactive design strategies. This revealed the need to follow nested design schemes for truly global and interactive optimizations that accurately considers model updating between design stages.

Conclusions: From this conceptual and numerical investigation, I summarize and conclude the following:

(1) Interactive design approaches allow to revise later data collection strategies based on new findings from earlier data. This approach allows the model to evolve during the DoE problem and the simultaneous data collection. This leads to an increased robustness and thus to a better performance of the data collection strategie.

(2) Two interactive strategies were investigated in a synthetic case study. Both led to a significantly higher overall performance compared to non-interactive strategies. In the best setup, it was possible to achieve the same performance compared to non-interactive DoE with only half of the measurements.

(3) Higher interaction frequencies generally lead to an improvement in the performance. The reduction of jointly optimized measurements in the sequential approach resulted in less complex designs, which tend to favor measurements providing high individual data impacts.

(4) The trade-off between earlier incorporation of data and ignorance of global optimization aspects in the sequential approach clearly favored higher feedback frequencies. The negative effect of possibly neglected redundancy effects or missing joint interactions between measurements are difficult to quantify. This is, of course, dependent on the investigated system, but clearly points out the importance of interaction versus the often discussed global optimization approaches.

(5) The adaptive approach, which follows a global optimization scheme, does not show an increased performance. I explained in the discussion that the approach is not able to globally optimize the expected data impact beyond an interactive Bayesian model update. It was pointed out that truly global optimization would require a recursively nested optimization loop. This will need further investigation to evaluate how far such recursive approaches would need to look ahead in order to foresee changing model states sufficiently well.

(6) In summary, the interactive design approach allows to use a faster and less complex optimization approach of sequentially optimizing ten single measurements. Despite of being faster and requiring only very basic search algorithms, it leads to the best and most robust overall performance.

7. Synergies

This chapter discusses shortly the synergy effects that arise from combining the findings of the previous three chapters. Chap. 4 revealed the importance of nonlinear and accurate data impact estimation for the design of data acquisition campaigns. Chap. 5 pointed out a significant speed-up potential by reversing the formulation of nonlinear data impact estimation, especially for low-dimensional designs. That chapter additionally introduced an approximated measure of data impact, which performed best in lower dimensions as well. At last, the previous Chap. 6 concluded that, when facing high uncertainty in the prior system model, sequential interactive design and data collection approaches performed superior to a global non-interactive approach. Thus lead to many subsequent low-dimensional design sequences that heavily favor reverse approaches for data impact analysis from Chap. 5. This small chapter offers a final integral view on the results and conclusions of this thesis and illustrates the potential benefits of their combination.

Vision of real-time design of data acquisition: Since the collection of data in the subsurface requires drilling or similar preparations, data acquisition in the subsurface often features longer interrupts between individual measurement collections. These interruptions between data acquisition can potentially be used for applying fast DoE methods for interactively designing the following measurements.

However, this requires that the DoE framework is sufficiently fast so that it can be applied within these time windows (e.g., deconstruction of drilling equipment or overnight), without delaying the field work. It could even be possible to evaluate this process on site, using a laptop or an installed workstation. This would allow to benefit from an interactive design and acquisition of data in a broad variety of applications in subsurface hydrology.

The focus of this chapter is therefore to evaluate the possibility of using the methods in this thesis for a real-time DoE framework. This would also require that the designs are evaluated on a (mobile) workstation, in contrast to previous scenarios, which relied on high-performance computation using more than 200 cores.

7.1. Combined nonlinear, reverse and interactive design of data acquisition

Each of the previous chapters dealt with a single facet of data impact estimation (accuracy, speed and robustness). A thoughtful combination of the reverse mindset for the utility evaluation within an interactive design framework may lead to additional synergy effects and allow

real-time applications. The foundation for this is the main conclusion of Chap. 6, which stated that the benefits of the interactive incorporation of available data in the design process outweigh the negative effects of splitting the high-dimensional design optimization problem in individual parts. Therefore, I consider the sequential approach (see Sec. 6.3.1) as the most beneficial approach for interactive framework. The combination of these findings introduces several synergy effects.

All reverse-based estimates of data impact from Chap. 5 showed the highest speed-up-potential for low-dimensional design problems (see Sec.). Additionally, approximation errors of linearized estimates are less severe, since only low-dimensional designs are evaluated and estimation errors do not propagate to later design sequences due to the fully nonlinear update after data acquisition. Finally, splitting the design in sequences and using approximations for data impact may allow to drastically lower the required ensemble sizes. When combining all these conclusions, the path to real-time interactive DoE of data acquisition is paved.

7.2. Application

Tab. 5.6 in the end of Chap. 5 provided a summary of all considered estimates of data impact, ranging from linear to nonlinear measures. I omit the entropy-based measures from this examination, since entropy-based data impact estimation requires large ensemble sizes and higher computational times, which hinders a real-time application on a mobile device.

Furthermore, I consider fast (and therefore approximated) estimates of data impact for this application: The first promising candidate is the reverse data impact based on variance $\Phi_{\text{rev var}}$ from Sec. 5.7. This method is partly linearized, requires minimal ensemble sizes and is fast. The second candidate is the EnKF-based approximation of data impact Φ_{EnKF} (see Sec. 3.3.1), which is fully linear. Both will compete against the fully nonlinear method of Chap. 4.

For additional speedup, I reduce the ensemble size such that an equivalent degree of statistical convergence is ensured for both implementations. A small preliminary convergence analysis for estimating data impact of a singe measurement is depicted in Fig. 7.1. It uses the bootstrapping method, that evaluates data impact for 50 repetitions to estimate the statistical error. The blue line shows the relative standard deviation (STD) of the EnKF-based utility function Φ_{EnKF} and the red line depicts the relative STD of the reverse approximated utility function $\Phi_{\text{rev var}}$. Both error measures fall below the defined error level of 5% (indicated by dotted black line) for MC-sizes larger than 10,000. Therefore, the MC-size in this analysis is set to 10,000 for both methods.

In order to compare the performance of the two considered real-time candidates to the previous fully nonlinear analysis of Chap. 6, I stay with the same application test case as introduced in chapter Sec. 6.5. The goal is to optimally design the collection of ten different measurements. The DoE problem is again performed for 50 synthetic realities to average over the influence of individual data sets.

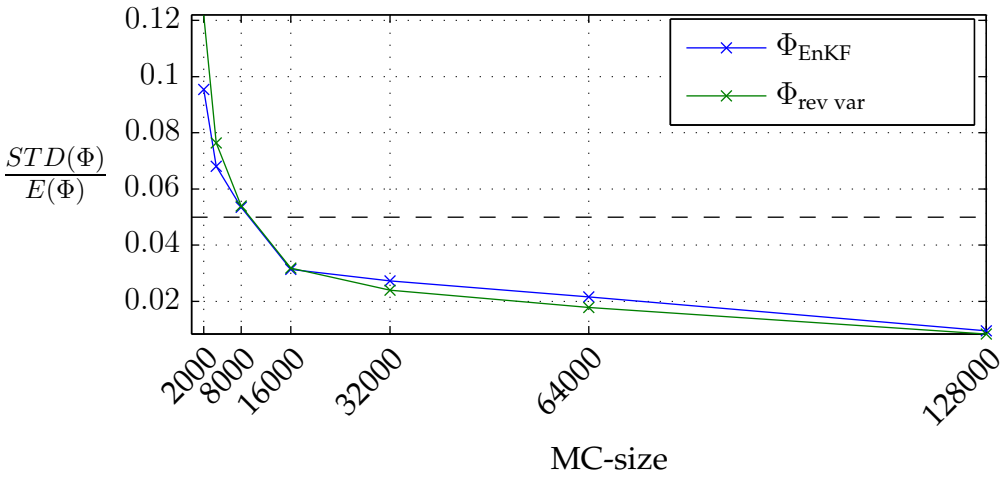


Figure 7.1.: Convergence analysis for the EnKF-based utility function and for the reverse approximation of data impact. Both implementations show similar convergence behavior and fall below an error level of 5 % for around 10.000 realizations.

7.3. Results

This section compares the results of the two real-time approaches with the previous fully nonlinear frameworks. I will first discuss the overall performance of the two used utility functions and then discuss briefly the achieved computation times for the design process.

The best results in the fully nonlinear test case are provided by the sequential approach (see Sec. 6.3.2), using the highest possible number of interaction sequences (see TC-B3 in Sec. 6.6.4). Therefore, I choose test case TC-B3 to serve as reference for comparison. This allows comparing the results of the real-time suitable solutions of this chapter with the performance of the best fully nonlinear approach.

7.3.1. Performance

Fig. 7.2 shows the overall performance, which is measured as the average posterior variance relative to the initial variance. The averaging is done over 50 synthetic realities and therefore indicates the general expected performance. The dotted black line shows the fully nonlinear, but non-interactive performance from Chap. 6 (global optimization). The red dotted line shows the fully nonlinear evaluation of data impact in combination with the sequential interaction approach from Sec. 6.3.2. Both are shown for the sake of comparison with the newly introduced approaches in this section.

Performance of the EnKF-based (linearized) method: The green line shows the final average performance of the EnKF-based linearized estimation of data impact. The superior per-

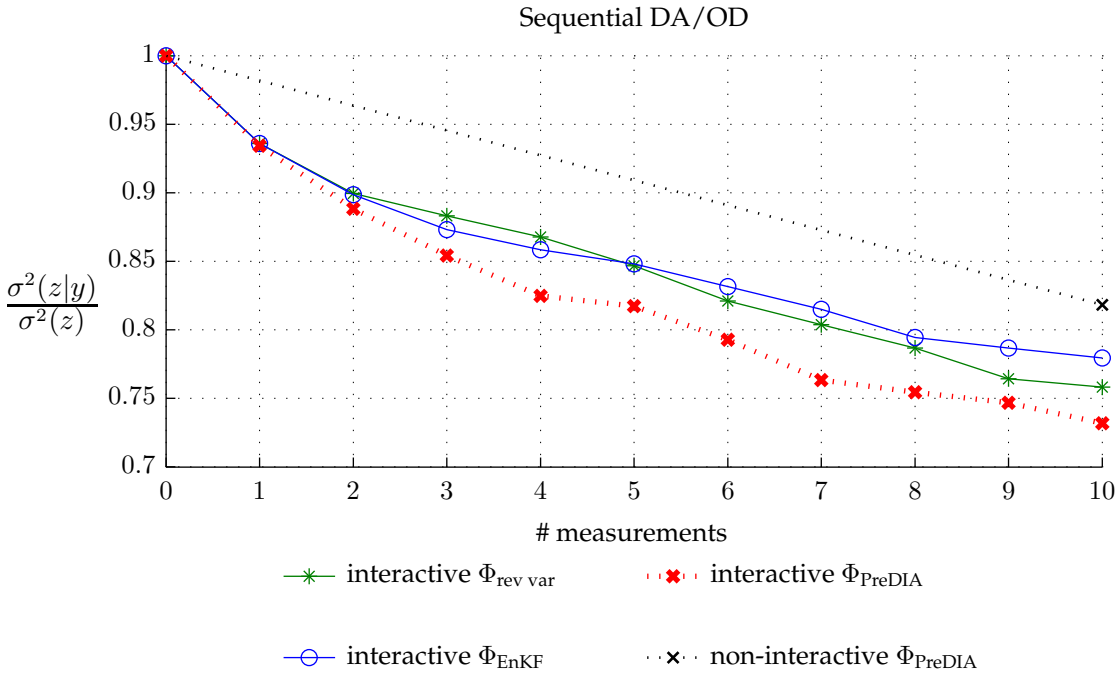


Figure 7.2.: Performance of real-time approaches of data acquisition, the EnKF-based in blue, the reverse approximation in green, and the fully nonlinear in red. All are compared to interactive and non-interactive fully nonlinear references in black.

formance compared to the non-interactive setup (dotted black line) indicates that sequential and fully nonlinear model updates using the real measurement data diminishes the estimation errors of the linear estimation: As the actual data impact is determined after each sequence, the estimation errors do not propagate to later design sequences and no linear data impact of multiple measurements is estimated jointly. In addition, the system still possesses a sufficiently large number of strongly linear dependencies, so that linear approximations are able to identify strong individual design candidates. Therefore, especially in the first sequences, the linear approach performs equally well compared to nonlinear estimates, since these designs sample the source zone, which is linearly dependent to the predicted concentration value (see Sec. 4.5.3).

Performance of the reverse variance-based method: The blue line shows the performance of the partly nonlinear approximation within the reverse mindset. The nonlinear evaluation of dependencies of measurement data to the prediction goal allows this approach to significantly outperform the linearized estimate: Compared to the EnKF-based approach, approximately the same uncertainty reduction can be achieved by eight measurements instead of ten. The better performance of the reverse approximation compared to the EnKF-based method can mostly be found in later design sequences. In these, the linear data has already been sampled and the remaining informative data are related nonlinearly to the prediction target. Still, the reverse approximation is inaccurate by assuming a Gaussian distribution for the measurements (see Sec. 5.7 . This leads to a performance reduction compared to the fully nonlinear evaluation,

depicted by the red dotted line in Fig. 7.2.

7.3.2. Evaluation times:

The evaluation times and used number of CPUs is summarized in Tab. 7.1. The non-interactive approach (black dotted line) required on average five hours for the evaluation using 40 CPUs. The genetic algorithm converged after 250 generations and evaluated a total of 25,000 designs.

In contrast, both real-time strategies were performed on a four-core workstation. The greedy search evaluated $\sim$ 1,400 designs per sequence, which results in only about 14,000 overall evaluated designs. The EnKF-based evaluation required 22 minutes for the design and update process of all ten sequences. The same evaluation using the reverse approximated data impact required 21 minutes.

In contrast, the fully nonlinear interactive approach required on average 25 hours using 40 CPUs using the same greedy search with about 15,000 overall evaluated designs.

Utility	interactive	number of sequences	measurements per sequence	design evaluations	number of cores	required time
Φ_{PreDIA}	no	1	10	$\sim$ 25,000	40	$\sim$ 6.5 h
Φ_{EnKF}	yes	10	1	$\sim$ 14,000	4	$\sim$ 0.37 h
$\Phi_{\text{rev var}}$	yes	10	1	$\sim$14,000	4	$\sim$ 0.35 h
Φ_{PreDIA}	yes	10	1	$\sim$ 15,000	40	$\sim$ 25 h

Table 7.1.: Summary of test cases and their key properties.

With this, solving the DoE problem is not the bottleneck of a real-time design framework anymore. The computational burden shifts to the generation of conditional realizations, based on the newly available measurement data after each data collection step. The time for this generation varies depending on the system complexity, dimensionality and the number and value of the available measurements.

The evaluation times for all approaches exclude the generation of conditional realizations as an already computed ensemble of realizations is used. How the generation of conditional realizations could be improved for the current context has to be investigated separately and calls for powerful and fast inversion techniques.

7.4. Discussion and conclusion

This chapter illustrated the advantages of a DoE framework that is specifically configured to match the complexity of the given system model and specific data acquisition campaign. Differently complex estimates of data impact allowed the modeler or engineer to balance the trade-off between optimization speed and estimation quality. This small synthetic study allows me to conclude the following:

1. Both approximations of data impact $\Phi_{\text{rev var}}$ and Φ_{EnKF} were sufficiently fast to be considered as real-time applicable, even on single computing units. The fully nonlinear and interactive approach Φ_{PreDIA} can be considered as well for real-time application, however requires distributed computing.
2. These efficient methods shift the computational bottleneck within interactive DoE away from the optimization towards the generation of conditional realizations.
3. The reverse approximation of data impact $\Phi_{\text{rev var}}$ performs significantly better than the fully linear approximation Φ_{EnKF}, while offering the same computational speed.
4. In each sequence, the collection of the actual measurements and their incorporation into the updated model replaces any prior estimate of data impact. Therefore, estimation errors of both approximations do not propagate to later sequences. As a result, linearized or partly linearized approaches perform significantly better within an interactive framework than within a non-interactive framework.
5. At last, splitting the full design task in multiple sub-designs allows for simple search methods. In particular, a simple greedy search is able to identify the global optimal design for one measurement.
6. However, both approximations were not able to perform as well as the fully nonlinear method (here: PreDIA from Chaps. 4-6). The disadvantage of PreDIA in the real-time context is that it requires significantly longer evaluation times.

8. Summary, conclusion and outlook

8.1. Summary

This thesis strived for efficient concepts within simulation-based optimal design of data acquisition for model calibration. For this context, efficiency of the overall optimization was broken down into three main facets: the estimation accuracy of data impact, evaluation speed of the overall optimization and estimation robustness related to uncertainties. The different facets and their respective questions were approached in three different main parts in this thesis, which are repeated in short:

Accuracy: The demand for accurate estimation of data impact in environmental systems led to the requirement to consider nonlinear statistical dependencies between data and prediction goals. In realistic applications potential data, model parameters and model prediction quantities are often connected nonlinearly. Furthermore, the various types of uncertainty in model parameters and concepts called for a flexibly applicable framework and led to the question:
Which available tools in the literature allow a flexible and accurate data impact estimation, how can they be combined most effectively and how can they be improved?

The first part (Chap. 4) identified the Bootstrap filter as the most suitable tool for fast nonlinear inversion. The method of PreDIA was developed as an extension of the Bootstrap filter towards a nonlinear data impact estimation technique. Combined with a brute-force Monte-Carlo approach, the framework is flexible to tackle complex design problems and consider any type of uncertainty.

Speed: The second goal was to speed up the expensive nonlinear data impact estimation such that complex and realistic design problems for data acquisition can be tackled. Evaluation speed is especially important within DoE frameworks, since during any optimization process, the data impact need to be estimated for hundreds of thousands to millions of designs proposals. Therefore, the basic information-theoretic measures to describe data impact have been reviewed to answer the question:
Which *theoretical* potential can be identified to accelerate nonlinear data impact estimation without using further approximations and how well can it be exploited in *practice*?

The second part (Chap. 5) found Mutual Information as the basic measure for information and thus for data impact. The symmetry of Mutual Information, and of Bayes Theorem respectively, provided the foundation for developing a reverse formulation of nonlinear data impact. This formulation equivalently evaluates the same Mutual Information, but at much faster evaluation times, by swapping the statistical roles of measurement data and model prediction. This new reverse mindset further allowed me to develop a partly linearized approximation of Mutual

Information, which provided superior estimates compared to linear techniques at comparable speed. A synthetic study showed a actual speedup in the range of one order of magnitude and more.

Robustness: The last considered facet of efficiency was the robustness/reliability of the evaluated estimates of data impact with respect to the uncertain prior model. A proper representation of the prior uncertain is required for an adequate assessment of the information needs and the simulation-based design process obviously depends on the prior belief state. This led to questions of how much this uncertainty affects the design process and its outcome. Thus, the most promising way to increase the estimation robustness was to incorporate new data as soon as possible in the design process. This led to the last research question:
What interactive mechanisms can be introduced to increase the robustness of nonlinear data impact estimation related to the uncertain system model and how much is the data impact improved by this interaction?

The third part of this thesis (Chap. 6) investigated the influence of the uncertain model on the subsequent design process for potential data acquisition. It highlighted the potential benefits of an interactive design and acquisition of measurements data, which is able to adapt the remaining measurement design to a constantly improving state of knowledge based on newly available measurement data. For the investigated system, the positive effects of earliest possible incorporation of data in the design process countered any negative effects that arose from a non-global optimization of the overall design.

Chap. 7 was highlighting the synergies between the three different steps above, when the particular design problem allows them to be combined efficiently. This last chapter depicted, how the interactive design approach can be used with the reverse estimation of data impact. In that context, it was shown that approximation errors in the data impact estimation have minor effects on the overall performance compared to their negative impact in non-interactive frameworks.

8.2. Summary of conclusions

Step I: Fully nonlinear and accurate assessment of data impact

This thesis developed a efficient framework for the assessment of nonlinear data impact that is generally applicable for various systems of any statistical dependency. The framework was tested for various applications. From a global point of view, the major conclusions were:

- Linear data impact estimates (e.g., EnKF-based, regression-like methods) fail to recognize relevant nonlinear relations between potential measurement locations and the prediction goal. Hence, linear methods oversample locations considered to be most informative from the limited viewpoint of linearized analysis. Fully nonlinear methods, (e.g., the PreDIA method developed here) overcome the shortcomings of linear methods and thus lead to superior designs in terms of actual data impact.

- The developed framework PreDIA can handle arbitrary task-driven formulations of optimal design and arbitrary types of uncertainties, which was demonstrated in multiple scenarios. Task-driven design formulations of data impact are important to focus on the information needs of the model directly related to the decision relevant quantities, which yield more specific and efficient design solutions.
- The developed nonlinear methods can handle BMA implicitly at no additional conceptual effort and allow tackling problems in the context of high model uncertainty. Including Bayesian model averaging into data acquisition design allows to reduce the subjectivity of prior assumptions.
- The flexibility of PreDIA was illustrated in examples from multiple engineering fields. This includes an application in the field of conceptual model choice, in which PreDIA was used to identify data, which has the highest potential to discriminate different structural and conceptual models.

Step II: Reverse formulation of nonlinear data impact

The second part of this thesis developed an equivalent, but significantly faster evaluation method for nonlinear data impact based on Mutual Information. The theoretic equivalence was proven and illustrated in different applications. Reversing data impact assessment further allowed to consider other approximations of data impact within the reverse mindset. One additional approximation was introduced to avoid the expensive *pdf*-estimation in high-dimensions that is required to work with MI. The most important findings and conclusions are:

- The reverse formulation for nonlinear data impact assessment is theoretically identical to the classical forward analysis. The numerical implementation of the reverse approach provides design quality values that are comparable to those of the forward formulation.
- The theoretical speedup potential leads to a practical speedup in most numerical test cases. The fastest implementation showed at least a speed-up of one order of magnitude and up to two orders of magnitude for small ensemble sizes and low numbers of measurements.
- The choice of the numerical technique for high-dimensional *pdf*-estimation is essential to optimally exploit the speedup promised by the theoretical considerations. In this context, GPU-based kernel density estimators performed best. The KNN-based entropy estimation performed best among the CPU-based estimators.
- In general, the new reverse approach allows modelers to apply nonlinear estimation of data impact within DoE for larger and more complex, realistic problems. The lean and less complex approach allows to choose from a broader variety of uncertainty estimation techniques.
- The reverse approximation of data impact based on the conditional multi-Gaussian assumption yields to tremendously shorter evaluation times and a much lower number of realizations. This approximation outperforms fully linear approaches, since it is able to assess parts of the nonlinear dependencies of the system.

Step III: Interactive coupling of Design of Experiments and data acquisition

The third part introduced the idea of interactive coupling between data acquisition design and its execution. The goal was to improve the robustness against uncertainty in the prior model and therefore to increase the overall performance. The evolving key conclusions from this chapter are:

- Most often for environmental models, a realistic representation of the current knowledge requires to choose the prior model very general in order to avoid unjustified assumptions. This high uncertainty in the available system model directly translates into high uncertainties in the assessment of the data impact, and hence reduces the reliability of optimally designed data acquisition campaigns.
- Sufficiently fast Design of Experiments frameworks allow to consider interactive design strategies that are able to adapt to the changing state of knowledge during the collection of data. This approach is more robust and outperforms globally optimized strategies, based solely on the prior (most uninformed) model belief.
- Introducing interactive strategies led in all test cases to a significantly higher overall performance compared to non-interactive strategies. In the investigated cases, the tested interactive approaches can achieve the same uncertainty reduction compared to non-interactive approaches with about 50 % of the number of measurements. The early incorporation of available data within the design process led to adapted designs that specifically follow the changing information needs of the evolving model.
- The trade-off between earlier incorporation of data and global en-block optimization in sequential approach clearly favored the earliest possible incorporation of data in sequential design. Relying on global optimization strategies did not lead to superior performance.
- Interactive design design strategies allow to drop the high-dimensional global optimization in favor of a less complex and faster sequential design approach, which nevertheless performs better. In applications in which an non-interactive, but global optimizations offers only slight improved design solutions, the usefulness of global optimization in an interactive setup is put into question.

Synergy effects

In a small concluding chapter, the combined use of the previous methods and insights allowed to benefit from the arising synergy effects. For particular problem classes, sacrificing some accuracy allows reducing the computational effort to evaluate an optimal data collection design so that much to allow its execution on a single workstation.

- Both approximations of data impact in an interactive setup were sufficiently fast to be considered as real-time applicable, even on single computing units. The fully nonlinear and interactive approach can be considered as well for real-time application, however requires distributed computing.

- The reverse approximation performed significantly better than the full linear method based on the EnKF. The combination of interactive design and the reverse mindset allows to quantify an approximated nonlinear data impact at the costs of linear analysis tools.
- All interactive approaches, including the fully linear, performed superior to the non-interactive nonlinear approach. This indicates: (i) Approximation errors are less significant within an interactive framework using subsequent low-dimension designs; (ii) The interactive model improvement and the consequential improvement in the data impact estimation is primarily affecting the overall performance.
- Relying on an interactive design approach and thus on a frequently changing state of knowledge shifts the computational bottleneck within interactive DoE away from the optimization towards the generation of conditional realizations. Therefore, fast simulation tools are required to generate the new ensemble-based representation of current state of knowledge.

8.3. Concluding recommendations

The optimal design of data acquisition for model calibration faces several challenges. The growing complexity of models and their rising amount of required calibration data call for effective design frameworks for optimized data acquisition. At the same time, the coupled and complex processes are modeled in more detail, which lead to higher nonlinear statistical dependencies. The results in this thesis clearly showed the importance of nonlinear estimation of data impact in environmental systems. For an efficient tackling of these challenges, I conclude the results of this thesis with the following recommendations:

First, the importance of the prior model on the design of data acquisition was outlined throughout the entire thesis. For a meaningful simulation-based optimal design, one needs to choose the prior model correctly including all types of uncertainties, including model choice. Only an accurately modeled prior state of knowledge allows for accurate estimates of information deficit and thus accurate data impact offered by the different design proposals. Only this allows to identify the actually relevant data.

The high computational effort of nonlinear data impact estimation does pay of for DoE of data acquisition. The combination with brute-force Monte-Carlo approaches is expensive, but allows to approach arbitrary classes of models and types of uncertainty. In contrast to the costs of data acquisition, the cost for computer power is constantly dropping.

Furthermore, the reverse mindset of data impact assessment shows that nonlinear data impact assessment are not necessarily expensive. The methods of this thesis massively reduced the computational costs for a larger range of task driven design problems. Yet, especially entropy estimation for arbitrary distribution in higher dimensions is still a challenge and therefore entropy-based information criteria are hardly used in applied engineering fields. Still, the methods presented in this thesis will benefit from the ongoing research for improved estimation techniques for high-dimensional distributions.

The best way to optimize and collect data for calibration of an uncertain simulation model is an interactive one. A local and adaptive design approach allows to efficiently focus on the current information needs and to adapt on the changing state of knowledge. By doing so, the model is not employed beyond its usefulness and accurately replace by the new model. The computational resources and time of a global foresightful optimization (based on a soon-to-be-obsolete model) is better invested to accurately represent the changing state of knowledge, as soon as new data becomes available.

Overall, the changing underlying model belief in nonlinear data impact estimation of future data and calls for specific solution strategies. Newly available data change the state of knowledge and therefore methods that do not adequately resemble these change of knowledge lack in their efficiency.

8.4. Outlook

Based on the conclusion above, this section highlights the related future research questions emerging from this work. Several **technical** issues can be identified for future research:

1. The method of PreDIA is based on two nested Monte-Carlo loops for the evaluation of nonlinear expectation. Variance reduction methods such as e.g., antithetic [*Hammersley and Morton*, 1956] or stratified sampling [e.g., *Särndal et al.*, 2003] may by used to speedup both loops. Therefore, future steps could investigate which variance reduction techniques serve best to effectively speed up nonlinear data impact analysis such as the PreDIA framework.

2. The introduced reverse formulation of data impact analysis does effectively speed up the data impact analysis, especially for a low number of measurements. For several measurements, the method relies heavily on high-dimensional density estimation techniques. Faster methods for estimation of high-dimensional *pdf*s, based on Voronoi elements [*Miller*, 2003] or spacing/partitioning approaches [e.g., *Stowell and Plumbley*, 2009] could increase the benefits of the reverse formulation even more.

3. To prove the applicability, the real-time capability of the design of data acquisition needs to be tested for the entire framework.

4. The generation of conditional realizations is is the new bottleneck in interactive DoE, when using the fast data impact analysis and the interactive approach from this thesis. Therefore, sufficiently fast methods for the generation of conditional realization need to be found. However, this general challenge needs to be solved specifically for any numerical model and depends highly on its complexity and required evaluation time.

In additional, on a **conceptual** level, future relevant research task are:

4. The partly linearized reverse approach is potentially fast enough to be used in a nested optimization. This would allow to elaborate designs that benefit from prospective anticipation of data interaction, but would suffer from increased approximation errors for the

data impact or from the same initial uncertainty. Therefore, future research needs to evaluate if this averaging over multiple event paths to incorporate learning effects into the design process actually leads to higher performance.

5. A conceptually simpler way to anticipate prospective learning is multi-objective optimization that jointly optimizes the expected value of data impact and the impact on variance of data impact of later sequences. This potentially identifies the designs with the highest impact on later design sequences among the designs with the highest expected data impact and allows to identify an adequate trade-off in a multi-objective optimization.

6. The conclusions that a global optimization beyond a model update is not useful needs to be tested in different physical scenarios, which differ in their initial uncertainty and in which a global optimization is specially important. This will help identifying cases in which the sequential data feedback suffices and in which the global approach generate better results.

7. Finally, optimizing data impact with respect to its expectation, indirectly gives credit to the reliability of the underlying statistical model description. However, admitting that the model is uncertain representation the real system, is might be better to base the optimization on more robust quantile of data impact, rather than on the expected value. This would acknowledge the possible erroneous and uncertain character of all model-based predictions, which includes model-based data impact.

Bibliography

Abellan, A., and B. Noetinger (2010), Optimizing subsurface field data acquisition using information theory, *Mathematical Geosciences*, *42*, 603–630, doi:10.1007/s11004-010-9285-6.

Abramowitz, M., and I. A. Stegun (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th ed., 1046 pp., Dover, New York.

Ammar, K., M. McKee, and J. Kaluarachchi (2011), Bayesian method for groundwater quality monitoring network analysis, *Journal of Water Resources Planning and Management*, *137*(1), 51–61, doi:10.1061/(ASCE)WR.1943-5452.0000043.

Arya, S., and D. M. Mount (1993), Approximate nearest neighbor queries in fixed dimensions, in *Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms*, SODA 93, pp. 271–280, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Ballio, F., and A. Guadagnini (2004), Convergence assessment of numerical Monte Carlo simulations in groundwater hydrology, *Water Resour. Res*, *40*(W04603), doi:10.1029/2003WR002876, W04603.

Bárdossy, A. (2006), Copula-based geostatistical models for groundwater quality parameters, *Water Resources Research*, *42*(11), W11,416, doi:10.1029/2005WR004754.

Bárdossy, A., and J. Li (2008), Geostatistical interpolation using copulas, *Water Resour. Res.*, *44*(W07412), doi:10.1029/2007WR006115.

Bear, J. (1972), *Dynamics of fluids in porous media*, 1st ed., 784 pp., Dover Publications, New York, USA.

Bear, J., and Y. Sun (1998), Optimization of pump-treat-inject (PTI) design for the remediation of a contaminated aquifer: multi-stage design with chance constraints, *Journal of Contaminant Hydrology*, *29*(3), 225–244, doi:10.1016/S0169-7722(97)00023-5.

Ben-Zvi, M., B. Berkowitz, and S. Kesler (1988), Pre-posterior analysis as a tool for data evaluation: application to aquifer contamination, *Water Resources Management*, *2*(1), 11–20, doi:10.1007/BF00421927.

Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrogeological Processes*, *6*(3), 279–298, doi:10.1002/hyp.3360060305.

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *Journal of Hydrology*, *249*, 11–29, doi:10.1016/S0022-1694(01)00421-8.

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

Bogaert, P., and D. Russo (1999), Optimal spatial sampling design for the estimation of the variogram based on a least squares approach, *Water Resources Research*, *35*(4), 1275–1289, doi: 10.1029/1998WR900078.

Box, G. E. P. (1982), Choice of response surface design and alphabetic optimality, *Utilitas Math.*, *21*, 11–55.

Brindle, A. (1981), Genetic algorithms for function optimisation, Ph.D. thesis, Department of Computing Science, University of Alberta.

Caflisch, R. (1998), Monte Carlo and quasi-Monte Carlo methods, *Acta numerica*, *7*, 1–49, doi: 10.1017/S0962492900002804.

Cerny, V. (1985), Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm, *Journal of Optimization Therory and Application*, *45*(1), 41–51, doi: 10.1007/BF00940812.

Chaloner, K., and I. Verdinelli (1995), Bayesian Experimental Design: A Review, *Statistical Science*, *10*(3), 273–304.

Chapman, S. (2008), Use of crop models to understand genotype by environment interactions for drought in real-world and simulated plant breeding trials, *Euphytica*, *161*(1-2), 195–208, doi:10.1007/s10681-007-9623-z.

Chen, Y., and D. Zhang (2006), Data assimilation for transient flow in geologic formations via ensemble Kalman filter, *Adv. Water Resour.*, *29*, 1107–1122, doi:10.1016/j.advwatres.2005.09.007.

Cheng, R. C. H. (1986), Variance Reduction Methods, in *Proceedings of the 18th Conference on Winter Simulation*, WSC '86.

Christakos, G. (1992), *Random field models in earth sciences*, 4th ed., 474 pp., Courier Dover Publications, New York.

Christakos, G., and B. R. Killam (1993), Sampling design for classifying contaminant level using annealing search algorithms, *Water Resources Research*, *29*(12), 4063–4076, doi:10.1029/93WR02301.

Christensen, S. (2004), A synthetic groundwater modelling study of the accuracy of GLUE uncertainty intervals, *Nordic hydrology*, *35*, 45–59.

Cirpka, O., and W. Nowak (2004), First-order variance of travel time in non-stationary formations, *Water Resources Research*, *40*(3), doi:10.1029/2003WR002851, W03507.

Cirpka, O. A., and P. K. Kitanidis (2001), Sensitivity of temporal moments calculated by the adjoint-state method and joint inversing of head and tracer data, *Adv. Water Resour.*, *24*(1), 89–103, doi:10.1016/S0309-1708(00)00007-5.

Cirpka, O. A., C. M. Bürger, W. Nowak, and M. Finkel (2004), Uncertainty and data worth analysis for the hydraulic design of funnel-and-gate systems in heterogeneous aquifers, *Water Resources Research*, *40*(11), doi:10.1029/2004WR003352, W11502.

Copty, N. K., and A. N. Findikakis (2000), Quantitative Estimates of the Uncertainty in the Evaluation of Ground Water Remediation Schemes, *Ground Water*, *38*(1), 29–37, doi:10.1111/j.1745-6584.2000.tb00199.x.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (1990), *Introduction to algorithms*, 1st ed., MIT press, USA.

Cover, T. M., and J. A. Thomas (2006), *Elements of Information Theory*, 2nd ed., 776 pp., Wiley-Interscience, Hoboken, New Jersey.

Criminisi, A., T. Tucciarelli, and G. P. Karatzas (1997), A methodology to determine optimal transmissivity measurement locations in groundwater quality management models with scarce field information, *Water Resour. Res.*, *33*(6), 1265–1274, doi:10.1029/97WR00300.

de Barros, F. P., D. Bolster, X. Sanchez-Vila, and W. Nowak (2011), A divide and conquer approach to cope with uncertainty, human health risk, and decision making in contaminant hydrology, *Water Resources Research*, *47*(5), doi:10.1029/2010WR009954, W05508.

de Barros, F. P., S. Ezzedine, and Y. Rubin (2012), Impact of hydrogeological data on measures of uncertainty, site characterization and environmental performance metrics, *Advances in Water Resources*, *36*(0), 51–63, doi:10.1016/j.advwatres.2011.05.004, special Issue on Uncertainty Quantification and Risk Assessment.

de Barros, F. P. J., and W. Nowak (2010), On the link between contaminant source release conditions and plume prediction uncertainty, *Journal of Contaminant Hydrology*, *116*(1–4), 24–34, doi:10.1016/j.jconhyd.2010.05.004.

de Barros, F. P. J., Y. Rubin, and R. M. Maxwell (2009), The concept of comparative information yield curves and its application to risk-based site characterization, *Water Resour. Res.*, *45*(6), doi:10.1029/2008WR007324, W06401.

Deb, K., and R. B. Agrawal (1994), Simulated binary crossover for continuous search space, *Tech. rep.*

Deutsch, C. V., and A. G. Journel (1997), *GSLIB: Geostatistical Software Library and Users Guide*, 2nd ed., 384 pp., Oxford University Press, New York.

Diggle, P., and S. Lophaven (2006), Bayesian geostatistical design, *Scandinavian Journal of Statistics*, *33*(3), 53–64, doi:10.1111/j.1467-9469.2005.00469.x.

Diggle, P. J., and P. J. Ribeiro (2007), *Model-based geostatistics*, Springer series in statistics, 1st ed., 230 pp., Springer, New York.

Diggle, P. J., and P. J. J. Ribeiro (2002), Bayesian inference in Gaussian model-based geostatistics, *Geographical and Environmental Modelling*, *6*(2), 129–146, doi:10.1080/1361593022000029467.

Doherty, J. (2002), *PEST: model independent parameter estimation* , 4th ed., Watermark Numerical Computing, Corinda, Australia.

Dougherty, D. E., and R. A. Marryott (1991), Optimal Groundwater Management: 1. Simulated Annealing, *Water Resources Research*, *27*(10), 2493–2508, doi:10.1029/91WR01468.

Edgeworth, F. Y. (1908), On the Probable Errors of Frequency-Constants, *Journal of the Royal Statistical Society*, *71*(3), pp. 499–512, doi:10.2307/2339378.

Engel, T., and E. Priesack (1993), Expert-N - A building block system of nitrogen models as resource for advice, research, water management and policy, in *Integrated soil and sedimentresearch: A basis for proper protection*, *Soil & Environment*, vol. 1, edited by H. Eijsackers and T. Hamers, pp. 503–507, Springer Netherlands.

Evensen, G. (1994), Inverse methods and data assimilation in nonlinear ocean models, *Physica D*, *77*, 108–129, doi:10.1016/0167-2789(94)90130-9.

Evensen, G. (2007), *Data Assimilation: The ensemble Kalman filter*, 2nd ed., 280 pp., Springer, Berlin, Heidelberg.

Federov, V., and P. Hackl (1997), *Model-Oriented Design of Experiments*, 1st ed., 117 pp., Springer-Verlag, New York, New York.

Fetter, C. W., and C. J. Fetter (1999), *Contaminant hydrogeology*, vol. 500, 2nd ed., 500 pp., Prentice Hall, Inc., Upper Saddle River, New Jersey.

Feyen, L. (2003), A Bayesian approach to stochastic capture zone delineation incorporating tracer arrival times, conductivity measurements, and hydraulic head observations, *Water Resources Research*, *39*(5), 1126, doi:10.1029/2002WR001544.

Feyen, L., and S. M. Gorelick (2005), Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas, *Water Resources Research*, *41*(3), doi:10.1029/2003WR002901, W03019.

Fishburn (1970), Utility theory for decision making, *Tech. rep.*

Fogg, G. E., C. D. Noyes, and S. F. Carle (1998), Geologically based model of heterogeneous hydraulic conductivity in an alluvial setting, *Hydrogeology Journal*, *6*(1), 131–143, doi:10.1007/s100400050139.

Ford, I., and D. Titterington (1989), Recent Advances in Nonlinear Experimental Design, *Technometrics*, *31*(1), 49–60, doi:10.2307/1270364.

Freeze, R., and S. Gorelick (1999), Convergence of Stochastic Optimization and Decision Analysis in the Engineering Design of Aquifer Remediation, *Water Resources Research*, *37*(6), doi:10.1111/j.1745-6584.1999.tb01193.x.

Freeze, R. A., B. James, J. Massmann, T. Sperling, and L. Smith (1992), Hydrogeological Decision Analysis: 4. The concept of data worth and its use in the development of site investigation strategies, *Ground Water*, *30*(4), 574–588, doi:10.1111/j.1745-6584.1992.tb01534.x.

Fritz, J., I. Neuweiler, and W. Nowak (2009), Application of FFT-based Algorithms for Large-Scale Universal Kriging Problems, *Mathematical Geosciences*, *41*(5), 509–533, doi:10.1007/s11004-009-9220-x.

Gates, J. S., and C. C. Kisiel (1974), Worth of additional data to a digital computer model of a groundwater basin, *Water Resour. Res*, *10*(5), 1031–1038, doi:10.1029/WR010i005p01031.

Gayler, S., E. Wang, E. Priesack, T. Schaaf, and F.-X. Maidl (2002), Modeling biomass growth, N-uptake and phenological development of potato crop, *Geoderma*, *105*(3–4), 367–383, doi:10.1016/S0016-7061(01)00113-6.

Genuchten, M. T. V. (1980), A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Science Society of America Journal*, *44*(5), 892–898.

Ghosh, S. (1999), *Multivariate analysis, design of experiments, and survey sampling*, Statistics: A Series of Textbooks and Monographs, Taylor & Francis, New York, USA.

Goldberg, D. E. (1989), *Genetic algorithms insearch, optimization and machine learning*, 1st ed., 432 pp., Addison-Wesley Longman, Boston, MA, USA.

Gomez-Hernandez, J. J., and X.-H. Wen (1998), To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology, *Adv. Water Resour.*, *21(1)*(1), 47–61, doi:10.1016/S0309-1708(96)00031-0.

Gómez-Hernández, J. J., A. Sahuquillo, and J. E. Capilla (1997), Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data - 1. Theory, *Journal of Hydrology*, *203*(1-4), 162–174.

Gopalakrishnan, G., B. Minsker, and D. Goldberg (2003), Optimal sampling in a noisy genetic algorithm for risk-based remediation design., *Journal of Hydroinformatics*, *5*, 11–25.

Gordon, N., D. Salmond, and A. Smith (1993), Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proceedings-F*, *140(2)*, 107–113.

Goudriaan, J., and H. V. Laar (1994), *Modelling potential crop growth processes: textbook with exercises*, vol. 2, Springer, USA.

Guest, T., and A. Curtis (2009), Iteratively constructive sequential design of experiments and surveys with nonlinear parameter-data relationships, *Journal of Geophysical Research: Solid Earth*, *114*(B4), doi:10.1029/2008JB005948.

Gueting, N., and A. Englert (2013), Hydraulic conditions at the source zone and their impact on plume behavior, *Hydrogeology Journal*, *21*(4), 829–844, doi:10.1007/s10040-013-0962-7.

Gunawan, S., and S. Azarm (2005), Multi-objective robust optimization using a sensitivity region concept, *Structural and Multidisciplinary Optimization*, *29*(1), 50–60, doi:10.1007/s00158-004-0450-8.

Guthke, P., and A. Bárdossy (2012), Reducing the number of MC runs with antithetic and common random fields, *Advances in Water Resources*, *43*(0), 1–13, doi:10.1016/j.advwatres.2012.03.014.

Haario, H., E. Saksman, J. Tamminen, et al. (2001), An adaptive Metropolis algorithm, *Bernoulli*, *7*(2), 223–242, doi:10.2307/3318737.

Haber, E., L. Horesh, and L. Tenorio (2008), Numerical methods for experimental design of large-scale linear ill-posed inverse problems, *Inverse Problems*, *24*(5), 055,012, doi:10.1088/0266-5611.

Hadka, D., and P. M. Reed (2013), Borg: An auto-adaptive many-objective evolutionary computing framework., *Evolutionary Computation*, *21*(2), 231–259, doi: 10.1162/EVCO_a_00075.

Hammersley, J. M., and K. W. Morton (1956), A new Monte Carlo technique: antithetic variates, *Mathematical Proceedings of the Cambridge Philosophical Society*, *52*, 449–475, doi:10.1017/S0305004100031455.

Handcock, M. S., and M. L. Stein (1993), A Bayesian analysis of kriging, *American Statistical Association and American Society for Quality*, *35(4)*(4), 403–410, doi:10.2307/1270273.

Hansen, N., and S. Kern (2004), Evaluating the CMA evolution strategy on multimodal test functions, in *Parallel Problem Solving from Nature - PPSN VIII*, *Lecture Notes in Computer Science*, vol. 3242, edited by X. Yao, E. Burke, J. Lozano, J. Smith, J. Merelo-Guervós, J. Bullinaria, J. Rowe, P. Tiňo, A. Kabán, and H.-P. Schwefel, pp. 282–291, Springer.

Harlow, F. H., and W. E. Pracht (1972), A theoretical study of geothermal energy extraction, *Journal of Geophysical Research*, *77*(35), 7038–7048, doi:10.1029/JB077i035p07038.

Herrera, G. S., and G. F. Pinder (2005), Space-time optimization of groundwater quality sampling networks, *Water Resources Research*, *41*, doi:10.1029/2004WR003626, W12407.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Statistical Science*, *14*(4), 382–417, doi:10.1214/ss.

Hou, Z., and Y. Rubin (2005), On minimum relative entropy concepts and prior compatibility issues in vadose zone inverse and forward modeling, *Water Resources Research*, *41*(12), n/a–n/a, doi:10.1029/2005WR004082.

Howard, R. A. (1966), Information value theory, *Systems Science and Cybernetics, IEEE Transactions on*, *2*(1), 22–26, doi:10.1109/TSSC.1966.300074.

Hu, H., J. Zhang, S. Guo, and G. Chen (1999), Extraction of Huadian oil shale with water in sub- and supercritical states, *Fuel*, *78*(6), 645–651, doi:10.1016/S0016-2361(98)00199-9.

Hu, I. (1998), On Sequential Designs in Nonlinear Problems, *Biometrika*, *85*(2), pp. 496–503, doi:10.1093/biomet.

James, B. R., and S. Gorelick (1994), When enough is enough: the worth of monitoring data in aquifer remediation design, *Water Resources Research*, *30*(12), 3499–3513, doi:10.1029/94WR01972.

Jaynes, E. T. (1957), Information theory and statistical mechanics, *The Physical Review*, *106*(4), 620–630, doi:10.1103/PhysRev.106.620.

Jong, K. A., and W. M. Spears (1991), An analysis of the interacting roles of population size and crossover in genetic algorithms, in *Parallel Problem Solving from Nature*, *Lecture Notes in Computer Science*, vol. 496, edited by H.-P. Schwefel and R. Männer, pp. 38–47, Springer Berlin Heidelberg, doi:10.1007/BFb0029729.

Journel, A., and C. Deutsch (1997), Rank order geostatistics: A proposal for a unique coding and common processing of diverse data, *Geostatistics Wollongong*, *96*(1), 174–187.

Journel, A. G., and C. J. Huijbregts (1978), *Mining Geostatistics*, Academic Press, New York.

Kass, R., and L. Wasserman (1996), The selection of prior distributions by formal rules, *Journal of the American Statistical Association*, *91*, 1343–1370, doi:10.2307/2291752.

Kennedy, J. F., J. Kennedy, and R. C. Eberhart (2001), *Swarm intelligence*, Morgan Kaufmann.

Kerrou, J., P. Renard, H. H. Franssen, and I. Lunati (2008), Issues in characterizing heterogeneity and connectivity in non-multiGaussian media, *Advances in Water Resources*, *31*(1), 147–159, doi:10.1016/j.advwatres.2007.07.002.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983), Optimization by simulated annealing, *Science*, *220*(4598), 671–680, doi:10.1126/science.220.4598.671.

Kitanidis, P. (1997), *Introduction to geostatistics: applications in hydrogeology*, Cambridge University Press.

Kitanidis, P. K. (1986), Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resources Research*, *22*(4), 499–507, doi:10.1029/WR022i004p00499.

Kitanidis, P. K. (1995), Quasi-Linear Geostatistical Theory for Inversing, *Water Resour. Res.*, *31*(10), 2411–2419, doi:10.1029/95WR01945.

Kitanidis, P. K. (1996), Analytical expressions of conditional mean, covariance, and sample functions in geostatistics, *Stochastic Hydrology and Hydraulics*, *10*(4), 279–294, doi:10.1007/BF01581870.

Kollat, J., P. Reed, and J. Kasprzyk (2008), A new epsilon-dominance hierarchical Bayesian optimization algorithm for large multiobjective monitoring network design problems, *Advances in Water Resources*, *31*(5), 828–845, doi:10.1016/j.advwatres.2008.01.017.

Kopp, A., H. Class, and R. Helmig (2009), Investigations on CO2 storage capacity in saline aquifers—Part 2: Estimation of storage capacity coefficients, *International Journal of Greenhouse Gas Control*, *3*(3), 277–287, doi:10.1016/j.ijggc.2008.10.001.

Krige, D. G. (1951), A statistical approach to some basic mine valuation problems on the Witwatersrand, *Journal of the Chemical, Metallugical and Mining Society of South Africa*, *52*, 119–139.

Kullback, S., and R. Leibler (1951), On information and sufficiency, *Annals of Mathematical Statistics*, *22*(1), 79–86.

Kunstmann, H., W. Kinzelbach, and T. Siegfried (2002), Conditional first-order second-moment method and its application to the quantification of uncertainty in groundwater modeling, *Water Resources Research*, *38(4)*(1035), 1035, doi:10.1029/2000WR000022.

Laplace, P. S. (1820), *Théorie analytique des probabilités*, Courcier.

LaVenue, A. M., B. S. RamaRao, G. D. Marsily, and M. G. Marietta (1995), Pilot Point Methodology for Automated Calibration of an Ensemble of Conditionally Simulated Transmissivity Fields 2. Application, *Water Resour. Res.*, *31*, 495–516, doi:10.1029/94WR02259.

Ledoux, M. (2001), *The concentration of measure phenomenon*, vol. 89, American Mathematical Society, Providence, USA.

Lepage, G. (1980), VEGAS-An adaptive multi-dimensional integration program, *Tech. rep.*, Research Report CLNS-80/447. Cornell University, Ithaca, N.-Y.

Leube, P., A. Geiges, and W. Nowak (2012), Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design, *Water Resources Research*, *48*(2), doi:10.1029/2010WR010137, W02501.

Liu, J. S. (2008), *Monte Carlo Strategies in Scientific Computing*, 360 pp., Springer, New York.

Liu, X., J. Lee, P. K. Kitanidis, J. Parker, and U. Kim (2012), Value of Information as a Context-Specific Measure of Uncertainty in Groundwater Remediation, *Water Resour. Res*, *26*(6), 1513–1535, doi:10.1007/s11269-011-9970-3.

Loaiciga, H. A. (1989), An optimization approach for groundwater quality monitoring network design, *Water Resources Research*, *25*(8), 1771–1782, doi:10.1029/WR025i008p01771.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, *330*(1-2), 368–381, doi:10.1016/j.jhydrol.2006.04.046.

Mariethoz, G., P. Renard, and J. Straubhaar (2010), The direct sampling method to perform multiple-point geostatistical simulations, *Water Resources Research*, *46*(11), doi:10.1029/2008WR007621, w11536.

Matérn, B. (1986), *Spatial variation*, 2nd ed., 151 pp., Springer, Berlin, Germany.

Matheron, G. (1971), *The Theory of Regionalized Variables and Its Applications*, Ecole de Mines, Fontainebleau, France.

Maxwell, R. M., W. E. Kastenberg, and Y. Rubin (1999), A methodology to integrate site characterization information into groundwater-driven health risk assessment, *Water Resources Research*, *35*(9), 2841–2855, doi:10.1029/1999WR900103.

Miller, E. (2003), A new class of entropy estimators for multi-dimensional densities, *Acoustics, Speech, and Signal Processing*, *3*, 297–300, doi:10.1109/ICASSP.2003.1199463.

Morariu, V. I., B. V. Srinivasan, V. C. Raykar, R. Duraiswami, and L. S. Davis (2008), Automatic online tuning for fast Gaussian summation, *Advances in Neural Information Processing Systems*, pp. 1113–1120.

Müller, W. G. (2007), *Collecting Spatial Data*, 3rd ed., Springer, Berlin, Germany.

Murakami, H., X. Chen, M. S. Hahn, Y. Liu, M. L. Rockhold, V. R. Vermeul, J. M. Zachara, and Y. Rubin (2010), Bayesian approach for three-dimensional aquifer characterization at the Hanford 300 Area, *Hydrology and Earth System Sciences*, *14*(10), 1989–2001, doi:10.5194/hess-14-1989-2010.

Myers, T. (2012), Potential contaminant pathways from hydraulically fractured shale to aquifers, *Groundwater*, *50*(6), 872–882, doi:10.1111/j.1745-6584.2012.00933.x.

Neuman, S., L. Xue, M. Ye, and D. Lu (2012), Bayesian Analysis of Data-Worth Considering Model and Parameter Uncertainties, *Adv. Water Resour.*, *36*, 75–85, doi:10.1016/j.advwatres.2011.02.007.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stoch. Environ. Res. Risk Assess.*, *17(5)*, 291–305, doi:10.1007/s00477-003-0151-7.

Neuman, S. P., and M. Ye (2009), Assessing and optimizing the worth of information under model, parameter and data uncertainties, *Eos Trans. AGU*.

Neupauer, R. M., and J. L. Wilson (2001), Adjoint-derived location and travel time probabilities for a multidimensional groundwater system, *Water Resources Research*, *37*(6), 1657–1668, doi:10.1029/2000WR900388.

Newman, M. E. J., and G. T. Barkema (1999), *Monte Carlo Methods in Statistical Physics*, Clarendon Press.

Nowak, W. (2008), A Hypothesis-Driven Approach to Site Investigation, *Eos Trans. AGU*, *89*(53), fall Meet. Suppl., Abstract H43A-0984.

Nowak, W. (2009a), Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator, *Water Resources Research*, *45*(4), W04,431, doi:10.1029/2008WR007328.

Nowak, W. (2009b), Measures of parameter uncertainty in geostatistical estimation and geostatistical optimal design, *Mathematical Geosciences*, *42*(2), 199–221, doi:10.1007/s11004-009-9245-1.

Nowak, W., and A. Litvinenko (2013), Kriging and spatial design accelerated by orders of magnitude: combining low-rank covariance approximations with FFT-techniques, *Mathematical Geosciences*, *45*(4), 411–435, doi:10.1007/s11004-013-9453-6.

Nowak, W., S. Tenkleve, and O. Cirpka (2003), Efficient computation of linearized cross-covariance and auto-covariance matrices of interdependent quantities, *Mathematical Geology*, *35*(1), 53–66, doi:10.1023/A:1022365112368.

Nowak, W., R. Schwede, O. Cirpka, and I. Neuweiler (2008), Probability density functions of hydraulic head and velocity in three-dimensional heterogeneous porous media, *Water Resources Research*, *44*(8), doi:10.1029/2007WR006383, W08452.

Nowak, W., F. P. J. de Barros, and Y. Rubin (2009), Bayesian geostatistical design - task-driven optimal site investigation when the geostatistical model is uncertain, *Water Resources Research*, *46*(3), 1944–7973, doi:10.1029/2009WR008312, W03535.

Nowak, W., Y. Rubin, and F. P. J. de Barros (2012), A hypothesis-driven approach to optimize field campaigns, *Water Resources Research*, *48*(6), doi:10.1029/2011WR011016, W06509.

Ogunsola, O. M., and N. Berkowitz (1995), Extraction of oil shales with sub- and near-critical water, *Fuel Processing Technology*, *45*(2), 95–107, doi:10.1016/0378-3820(95)00036-7.

Olivella, S., J. Carrera, A. Gens, and E. Alonso (1994), Nonisothermal multiphase flow of brine and gas through saline media, *Transport in Porous Media*, *15*(3), 271–293, doi:10.1007/BF00613282.

Oreskes, N., K. Shrader-Frechette, and K. Belitz (1994), Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, *263*(5147), 641–646, doi:10.1126/science.263.5147.641.

Pappenberger, F., and K. J. Beven (2006), Ignorance is bliss: Or seven reasons not to use uncertainty analysis, *Water Resour. Res.*, *42*(W05302), doi:10.1029/2005WR004,820.

Plackett, R. L., and J. P. Burman (1946), The design of optimum multifactorial experiments, *Biometrika*, *33*(4), 305–325, doi:10.2307/2332195.

Poeter, E. P., and M. C. Hill (1998), *Documentation of UCODE: A computer code for universal inverse modeling*.

Priesack, E., and S. Gayler (2009), Agricultural crop models: concepts of resource acquisition and assimilate partitioning, *Progress in Botany*, *70*(195), doi:10.1007/978-3-540-68421-3_9.

Priesack, E., S. Gayler, and H. Hartmann (2007), The impact of crop growth model choice on the simulated water and nitrogen balances, in *Modelling water and nutrient dynamics in soil–crop systems*, edited by K. Kersebaum, J.-M. Hecker, W. Mirschel, and M. Wegehenkel, pp. 183–195, Springer Netherlands, doi:10.1007/978-1-4020-4479-3_13.

Pukelsheim, F. (2006), *Optimal Design of Experiments*, Classics in Applied Mathematics, SIAM, Philadelphia, USA.

RamaRao, B. S., A. M. LaVenue, G. de Marsily, and M. G. Marietta (1995), Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: 1. Theory and computational experiments, *Water Resour. Res.*, *31(3)*, 475–493, doi:10.1029/94WR02258.

Reed, P., B. Minsker, and D. E. Goldberg (2000), Designing a competent simple genetic algorithm for search and optimization, *Water Resources Research*, *36*(12), 3757–3762, doi:10.1029/2000WR900231.

Reed, P. M., D. Hadka, J. D. Herman, J. R. Kasprzyk, and J. B. Kollat (2013), Evolutionary multiobjective optimization in water resources: The past, present, and future, *Advances in Water Resources*, *51*, 438–456, doi:10.1016/j.advwatres.2012.01.005.

Richards, L. A. (1931), Capillary conduction of liquids through porous mediums, *Journal of Applied Physics*, *1*(5), 318–333, doi:10.1063/1.1745010.

Ritchie, J., D. Godwin, and S. Otter-Nacke (1985), CERES-Wheat. A simulation model of wheat growth and development, *ARS*, pp. 159–175.

Rouhani, S. (1985), Variance Reduction Analysis, *Water Resources Research*, *21*(6), 837–846, doi:10.1029/WR021i006p00837.

Rubin, Y. (2003), *Applied stochastic hydrogeology*, 1st ed., 416 pp., Oxford University Press, USA, New York.

Russell, C. H. C. (1986), Variance Reduction Methods, in *Proceedings of the 18th conference on Winter simulation*, pp. 60–68, doi:10.1145/318242.318261.

Sahinidis, N. V. (2004), Optimization under uncertainty: state-of-the-art and opportunities, *Computers & Chemical Engineering*, *28*(6–7), 971–983, doi:10.1016/j.compchemeng.2003.09.017, {FOCAPO} 2003 Special issue.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008), *Global sensitivity analysis: the primer*, Wiley, New York.

Särndal, C.-E., B. Swensson, and J. Wretman (2003), *Model assisted survey sampling*, Springer, New York.

Scheidegger, A. E. (1961), General Theory of dispersion in porous media, *J. Geophys. Res.*, *66*(10), 3273–3278, doi:10.1029/JZ066i010p03273.

Schöninger, A., W. Nowak, and H.-J. Franssen (2012), Parameter estimation by Ensemble Kalman Filters with transformed data: Approach and application to hydraulic tomography, *Water Resour. Res.*, *48*, doi:10.1029/2011WR010462, w04502.

Schwede, R. L., O. A. Cirpka, W. Nowak, and I. Neuweiler (2008), Impact of sampling volume on the probability density function of steady state concentration, *Water Resources Research*, *44*(W12433), doi:10.1029/2007WR006668, W12433.

Schweppe, F. C. (1973), *Uncertain Dynamic Systems*, 1st ed., 576 pp., Prentice-Hall, Englewood Cliffs, NJ.

Scott, D. W. (2008), *Multivariate Density Estimation*, John Wiley & Sons, Inc., New York, doi:10.1002/9780470316849.

Shapiro, A. (2008), Stochastic programming approach to optimization under uncertainty, *Mathematical Programming*, *112*(1), 183–220, doi:10.1007/s10107-006-0090-4.

Shapiro, A., and T. Homem-de Mello (1998), A simulation-based approach to two-stage stochastic programming with recourse, *Mathematical Programming*, *81*(3), 301–325, doi:10.1007/BF01580086.

Shapiro, A., D. Dentcheva, et al. (2009), *Lectures on stochastic programming: modeling and theory*, vol. 9, SIAM Publications.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, 1st ed., 176 pp., Chapman & Hall/CRC, London, New York.

Simunek, J., M. T. V. Genuchten, and M. Sejna (2005), The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media, *University of California-Riverside Research Reports*, *3*, 1–240.

Singh, H., N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk (2003), Nearest Neighbor Estimates of Entropy, *American Journal of Mathematical and Management Sciences*, *23*(3-4), 301–321, doi:10.1080/01966324.2003.10737616.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson (2008), Obstacles to high-dimensional particle filtering, *Monthly Weather Review*, *136(12)*(12), 4629–4640, doi:10.1175/2008MWR2529.1.

Srinivasan, B., and Q. R. Duraiswami (2010), GPUML: Graphical processors for speeding up kernel machines, *Siam Conference on Data Mining, Workshop on High Performance Analytics - Algorithms, Implementations, and Applications*.

Storn, R., and K. Price (1997), Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces, *Journal of global optimization*, *11*(4), 341–359, doi:10.1023/A:1008202821328.

Stowell, D., and M. Plumbley (2009), Fast Multidimensional Entropy Estimation by -d Partitioning, *Signal Processing Letters, IEEE*, *16*(6), 537–540, doi:10.1109/LSP.2009.2017346.

Sun, N.-Z. (1994), *Inverse problems in groundwater modeling*, 1st ed., 352 pp., Springer, Dordrecht, Netherland.

Sun, N.-Z., and W. W.-G. Yeh (1990), Coupled inverse problems in groundwater modeling. 1. Sensitivity analysis and parameter identification, *Water Resources Research*, *26*(10), 2507–2525, doi:10.1029/WR026i010p02507.

Sykes, J. F., J. L. Wilson, and R. W. Andrews (1985), Sensitivity analysis for steady-state groundwater-flow using adjoint operators, *Water Resources Research*, *21*(3), 359–371, doi:10.1029/WR021i003p00359.

Tarantola, A. (2005), *Inverse Problem Theory*, SIAM Publications.

Tung, Y. (1986), Groundwater Management by ChanceConstrained Model, *Journal of Water Resources Planning and Management*, *112*(1), 1–19, doi:10.1061/(ASCE)0733-9496(1986)112:1(1).

Van Leeuwen, P. J. (2009), Particle filtering in geophysical systems, *Monthly Weather Review*, *137*(12), 4089–4114, doi:10.1175/2009MWR2835.1.

Vargas-Guzmán, J. A., and T.-C. J. Yeh (2002), The successive linear estimator: a revisit, *Adv. Water Resour.*, *25*(7), 773–781, doi:10.1016/S0309-1708(02)00066-0.

Vrugt, J. A., C. J. ter Braak, H. V. Gupta, and B. A. Robinson (2008), Equifinality of formal (DREAM) and Informal (GLUE) Bayesian Approaches in Hydrologic Modeling?, *Stoch. Environ. Res. Risk Assess.*, *23*, 1011–1026, doi:10.1007/s00477-008-0274-y.

Walter, L., P. J. Binning, S. Oladyshkin, B. Flemisch, and H. Class (2012), Brine migration resulting from {CO2} injection into saline aquifers – An approach to risk estimation including various levels of uncertainty, *International Journal of Greenhouse Gas Control*, *9*(0), 495–506, doi:10.1016/j.ijggc.2012.05.004.

Wand, M. P., and M. C. Jones (1995), *Kernel Smoothing*, 1st ed., 224 pp., CRC Press, Florida, USA.

Wang, E. (1997), *Development of a generic process-oriented model for simulation of crop growth*, Herbert Utz Verlag.

Weiss, N. A. (2006), *A Course in Probability*, 1st ed., 816 pp., Addison Wesley Longman, USA.

Wöhling, T., A. Geiges, W. Nowak, S. Gayler, P. Högy, and H. Wizemann (2013), Towards optimizing experiments for maximum-confidence model selection between different soil-plant models, *Procedia Environmental Sciences*, *19*(0), 514–523, doi:10.1016/j.proenv.2013.06.058.

Wolpert, D., and W. Macready (1997), No free lunch theorems for optimization, *Evolutionary Computation, IEEE Transactions on*, *1*(1), 67–82, doi:10.1109/4235.585893.

Woodbury, A. D., and T. J. Ulrych (1993), Minimum relative entropy: Forward probabilistic modeling, *Water Resour. Res.*, *29*(8), 2847–2860, doi:10.1029/93WR00923.

Yang, C., R. Duraiswami, N. Gumerov, and L. Davis (2003), Improved fast Gauss transform and efficient kernel density estimation, pp. 464–471, IEEE International Conference on Computer Vision, Nice, France, doi:10.1109/ICCV.2003.1238383.

Yeh, T.-C. J., M. Jin, and S. Hanna (1996), An iterative stochastic inverse method: conditional effective transmissivity and hydraulic head fields, *Water Resour. Res.*, *32*(1), 85–92.

Yokota, F., and K. M. Thompson (2004), Value of information analysis in environmental health risk management decisions: past, present, and future, *Risk Analysis*, *24*(3), 635–650, doi:10.1111/j.0272-4332.2004.00464.x.

Zang, C., M. Friswell, and J. Mottershead (2005), A review of robust optimal design and its application in dynamics, *Computers & Structures*, *83*(4), 315–326, doi:10.1016/j.compstruc.2004.10.007.

Zhang, Y., G. F. Pinder, and G. S. Herrera (2005), Least cost design of groundwater quality monitoring networks, *Water Resour. Res.*, *41(8)*, doi:10.1029/2005WR003936, W08412.

Zinn, B., and C. F. Harvey (2003), When good statistical models of aquifer heterogeneity go bad: A comparison of flow, dispersion, and mass transfer in connected and multivariate Gaussian hydraulic conductivity fields, *Water Resources Research*, *39(3)*(1051), 1051, doi:10.1029/2001WR001146.

A. Appendix: Gaussian summation

The evaluation of Gaussian sums in high dimensions is an essential element in both the forward and the reverse analysis. It is a main element in likelihood estimation in the bootstrap filter (see Sec. 5.4.2) and in any kernel-based *pdf* estimation (see Sec. 2.2.3), and requires the main computation time in the problems featured here. In general, the sum

$$g(\mathbf{y}_j) = \sum_{i=1}^{N} q_i e^{-(\mathbf{x}_i - \mathbf{y}_j)^2/h^2} \qquad j = 1...M \tag{A.1}$$

is computed, where $\mathbf{x}_1...\mathbf{x}_N$ and $\mathbf{y}_1...\mathbf{y}_M$ are source and target points, respectively, i.e., sample values within the ensemble and points of query in the *pdf* or likelihood estimation problem. q_i is an optional (situation-specific) weighting for individual target points $\mathbf{x}_i$ and h defines the bandwidth of the Gaussian Kernel function. An important sub-problem is the distance matrix $\mathbf{D}$ with

$$d_{ij} = -(\mathbf{x}_i - \mathbf{y}_j)^2/h^2 \tag{A.2}$$

that defines the distance between all source and target points in the high-dimensional space, normalized by the bandwidth parameter h. After $\mathbf{D}$ is known, exponentiation and summation over i yield the estimate $g(\mathbf{y}_j)$. The length of the vectors $\mathbf{x}_i$ and $\mathbf{y}_j$ corresponds to the number of points in the design, and thus corresponds to the dimensionality of the problem. Three possible numerical implementations are used throughout this study.

- The first option is a straightforward and vectorized implementation of Eqs. (A.1)-(A.2) in MATLAB.
- The second implementation is the C++ library FigTree [*Morariu et al.*, 2008], which is freely available under the Lesser General Public License (LGPL). FigTree is short for 'Fast Improved Gauss Transform with Tree Data Structure' and is designed to efficiently compute Gaussian sums in multiple dimensions, using the Improved Fast Gauss Transform [*Yang et al.*, 2003]. It also uses Approximate Nearest Neighbor searching [*Arya and Mount*, 1993], clustering of nearby source and target points and truncation of the Gaussian function for large distances to speed up the calculation of the distance matrix $\mathbf{D}$. The overall final accuracy of FigTree is controlled by a user-set error limit. More details can be found in [*Morariu et al.*, 2008].
- The third implementation is a CUDA-based library called GPUML (Graphical Processing Units for Machine Learning), see *Srinivasan and Duraiswami* [2010]. The library contains no approximation, but achieves its speed-up by outsourcing the summation to the computer's Graphical Processing Unit (GPU). The GPU is a massively parallelized computation unit, which perfectly suits for parallelized summation of Gaussian functions.

B. Appendix: Used computational hardware

All evaluation times were obtained on the high performance computing cluster 'Symphony', which was financed by the Stuttgart Research Center for Simulation Technology and the cluster 'Simulation Technology'. It was manufactured by the Dutch manufacturer 'Cluster Vision' and consists of 36 individual nodes.

When referred to CPU architecture, a single node is equipped with two AMD Quad-Core Opteron processors, 800 Mhz, 32 GB memory was used.

GPU-based evaluation was done on the included 'Nvidia GeForce GTX 285' graphic cards, using the CUDA language. This was done by the MATLAB-CUDA interface provided by *Srinivasan and Duraiswami* [2010].

Institut für Wasser- und Umweltsystemmodellierung
Universität Stuttgart

Pfaffenwaldring 61
70569 Stuttgart (Vaihingen)
Telefon (0711) 685 - 64717/64749/64752/64679
Telefax (0711) 685 - 67020 o. 64746 o. 64681
E-Mail: iws@iws.uni-stuttgart.de
http://www.iws.uni-stuttgart.de

Verzeichnis der Mitteilungshefte

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

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

3 Gurr, Siegfried: *Beitrag zur Berechnung zusammengesetzter ebener Flächentragwerke unter besonderer Berücksichtigung ebener Stauwände, mit Hilfe von Randwert- und Lastwertmatrizen*, 1965

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

97 Betz, Christoph: *Wasserdampfdestillation von Schadstoffen im porösen Medium: Entwicklung einer thermischen In-situ-Sanierungstechnologie*, 1998, ISBN 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: Upscaling, 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, 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 Hochwasserrückhaltebecken-systemen,* 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 3-933761-57-3

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

Die Mitteilungshefte ab der Nr. 134 (Jg. 2005) stehen als pdf-Datei über die Homepage des Instituts: www.iws.uni-stuttgart.de zur Verfügung.