Ensemble Kalman filtering for parameter estimation in groundwater flow modeling: implementation and robust comparison of filter variants within the SHEMAT-Suite Stochastic Mode

Von der Fakultät für Georessourcen und Materialtechnik der
Rheinisch-Westfälischen Technischen Hochschule Aachen

zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigte Dissertation

vorgelegt von

Johannes Joachim Keller, Master of Science

aus Würzburg

Berichter: Univ.-Prof. Dr. rer. nat. Harrie-Jan Hendricks Franssen
Univ.-Prof. Dr. rer. nat. Christoph Clauser
Univ.-Prof. Dr. sc. Julia Kowalski
Univ.-Prof. Dr.-Ing. Wolfgang Nowak

Tag der mündlichen Prüfung: 08.12.2020

Contents

SUMMARY

Introduction Modeling groundwater flow is important for many scientific and commercial applications related to the geosciences. Two such applications are the simulation of geothermal systems where a combination of flow and heat is simulated, and monitoring contaminant transport where a combination of flow and species transport is simulated. All such applications have in common that uncertainty quantification is essential, in particular the correct modeling of permeabilities of the porous medium through which the fluid, often water, flows. This thesis presents a framework for applying the ensemble Kalman filter (EnKF) for permeability estimation. The EnKF is a method for data assimilation and parameter estimation adapted to large-scale, non-linear models. In the first part of the thesis, the focus is on the numerical code SHEMAT-Suite.

SHEMAT-Suite is described, in particular the workflow for ensuring error-free and reproducible implementation that was implemented during this thesis. This concerns two parts of the software: the groundwater flow simulation and the EnKF update. A special focus is laid on the stochastic mode of SHEMAT-Suite. During this thesis, the existing software that allowed the usage of a potentially damped EnKF was completely refactored. This included among other tasks: a module implementation of the methods, a remodeling of the input file of the stochastic mode to match the main input file of the software, and, finally, including numerous error messages. Additionally, the suite of EnKF methods that is used in the remainder of this thesis was implemented in the same modular way. In the future, the modular implementation facilitates adding EnKF methods and robustly comparing the existing EnKF methods.

In the remainder of this thesis, I use this framework for investigating possible inaccuracies related to the EnKF and groundwater flow simulation. First, the influence of the sampling error stemming from random seed initialization is analyzed in a comparison of EnKF variants. Secondly, the pilot point EnKF is introduced. The pilot point EnKF is a novel EnKF method that is good at suppressing unwanted spurious correlations that may occur when using EnKF methods for parameter estimation.

SHEMAT-Suite In Chapter 3 of this thesis, attention is paid to the implementation of the scientific software SHEMAT-Suite. SHEMAT-Suite is an open-source numerical code for simulating flow, heat and mass transport in porous media. Additionally, SHEMAT-Suite can be used for parameter estimation and data assimilation by stochastic approaches (Monte Carlo, ensemble Kalman filter). The scientific impact of SHEMAT-Suite is illustrated and examples for using the software are presented. The following two scientific software developments implemented during the work on this thesis are presented: the modular implementation of a suite of EnKF methods, and using the software git for refactoring SHEMAT-Suite into branches optimized for its diverse functionalities. Afterwards, it is explained how the implementation of the EnKF methods and the refactoring using git were motivated by the scientific principles (1) reproducibility, (2) falsifiability, (3) communication of results, and (4) a sound foundation for future work. I conclude

the chapter by proposing guidelines for scientific software development that were derived from the development of SHEMAT-Suite.

Comparison of EnKF methods Chapter 4 of this thesis is dedicated to studying the influence of random seed variations on permeability estimation using the EnKF. Principally, all ensemble-based parameter estimation is influenced by the choice of random seed. Usually, the ensemble size of the parameter estimation method can be chosen so large that the influence of the random seed becomes negligible. However, for large-scale subsurface models, one is usually restricted to a small ensemble size. In this case, sampling fluctuations often render the influence of the random seed non-negligible. In Chapter 4, I argue that the knowledge of the random sampling error is essential for comparing the performance of multiple EnKF variants. To support this argument, the following large number of performance comparisons is carried out: for seven ensemble sizes (50, 70, 100, 250, 500, 1 000, 2 000) and seven EnKF variants (damped, iterative, local, hybrid, dual, normal score and classical EnKF), 1 000 synthetic parameter estimation experiments are computed for two setups: a 2-D tracer transport problem and a 2-D flow problem with one injection well. Up to now, multiple synthetic experiments with varied random seed are seldom shown in the literature, when the performance of EnKF methods is compared. By letting the synthetic parameter estimations differ only in random seed, 1 000 estimated permeability fields are collected and compared to a synthetic true permeability field using the root-mean-square error (RMSE). From comparing mean RMSEs for the seven EnKF methods and ensemble size 50, I conclude that a minimum of 10 synthetic experiments are needed to correctly distinguish RMSE differences smaller than 10 % between EnKF variants. For detecting RMSE differences smaller than 2 %, as many as 100 synthetic experiments are needed for the small, but commonly used, ensemble sizes 50, 70, 100, and 250.

Pilot point ensemble Kalman filter In Chapter 5 of this thesis, the pilot point ensemble Kalman filter (PP-EnKF) is presented, a variant of the ensemble Kalman filter that is tailored towards suppressing spurious correlations. Spurious correlations may bias parameter estimation results of EnKF methods when small ensemble sizes are used. Making use of the results from the first part of the thesis, the PP-EnKF is compared to the seven other EnKF methods using the existing comparison framework. For the 2-D tracer transport problem, the PP-EnKF ranks third out of eight methods in an overall performance ranking. For the 2-D flow problem and ensemble size 50, the PP-EnKF yields overall standard deviations that are closer to a reference than the overall standard deviations of all other EnKF methods. Additionally, the locally distributed correlations between observations and updated parameters are significantly closer to a reference for the PP-EnKF than for the classical EnKF. This highlights the method's capability for suppressing spurious correlations in parameter estimation using small ensemble sizes. On the basis of the results from this comparison, I promote the PP-EnKF for usage in permeability estimation studies.

Conclusion The results of this thesis provide insights for the robust and efficient usage of growing computer resources for permeability estimation and groundwater flow estimation. In the majority of scientific disciplines, quantifying uncertainties is of high importance and, due to improving computer performance, uncertainty quantification becomes feasible for previously too computationally intensive simulations. For permeability estimation and groundwater flow simulation, quantifying uncertainty is particularly essential for a number of reasons. (1) There are usually many uncertain influences in subsurface models used for groundwater flow, (2) subsurface models are typically large-scale, and, (3) as a consequence, small ensemble sizes have to suffice for ensemble-based uncertainty quantification methods. In this work, a framework for using ensemble Kalman filters (EnKF) for parameter estimation is presented. In this framework, the EnKF methods are implemented as part of the scientific software SHEMAT-Suite in such a way that the implementation supports scientific principles, such as reproducibility and falsifiability. By comparing performances of EnKF methods, it is shown that the random seed has a non-negligible influence on these performance comparisons. Finally, the PP-EnKF is introduced, an EnKF variant tailored towards suppressing spurious correlations that result from undersampling in the EnKF.

ZUSAMMENFASSUNG

Einleitung Die Modellierung von Grundwasserströmung ist ein wichtiger Aspekt vieler Anwendungsbereiche der Geowissenschaften. Ein Beispiel eines solchen Anwendungsbereichs ist die Modellierung und numerische Simulation geothermischer Systeme. Bei einer numerischen Simulation im geothermischen Anwendungsbereich werden Grundwasserströmung und Wärmeausbreitung gekoppelt berechnet. Ein weiteres Anwendungsbeispiel ist die Überwachung von Schadstofftransport in Grundwasserleitern. In diesem Fall werden Grundwasserströmung und Massentransport gekoppelt berechnet. Damit die Modellierung einer Grundwasserströmung möglichst aussagekräftig ist, bedarf es der Unsicherheitsquantifizierung, insbesondere der Kenntnis der Permeabilitäten des porösen Mediums, durch welches die Flüssigkeit, in den meisten Fällen Wasser, fließt. In dieser Doktorarbeit wird ein Softwarepaket vorgestellt, welches die Anwendung des Ensemble Kalman Filters (EnKF) für die Permeabilitätsschätzung ermöglicht. Der EnKF ist eine Methode zur Datenassimilierung und Parameterschätzung in rechenintensiven, nicht-linearen numerischen Simulationen.

Am Anfang dieser Doktorarbeit liegt der Fokus auf der numerischen Software SHEMAT-Suite, welche in Kapitel 3 beschrieben wird. In diesem Zusammenhang wird erläutert, durch welche Maßnahmen aus dieser Arbeit eine Fehlerreduktion bei der Softwareentwicklung und gleichzeitig eine möglichst einfach nachvollziehbare Softwareentwicklung gewährleistet werden kann. Von besonderer Wichtigkeit sind hierbei die folgenden zwei Bestandteile von SHEMAT-Suite: Die Grundwasserströmungssimulation und die Korrekturgleichung des EnKF. Desweiteren wird der Teil der Software SHEMAT-Suite besprochen, der stochastische Rechnungen ermöglicht. Im Rahmen dieser Arbeit ist die bestehende Software, die einen potentiell gedämpften EnKF beinhaltete, komplett überarbeitet worden. Folgende Punkte gehören zu dieser Überarbeitung: die Implementierung wurde modular gegliedert, die Eingabedatei wurde erweitert und in ihrer Form der Haupteingabedatei angepasst, zahlreiche Fehlerausgaben wurden eingeführt. Schließlich wurden die restlichen EnKF-Methoden, die in dieser Arbeit ein Rolle spielen, implementiert. Durch die Modularität der Implementierung ist es in Zukunft möglich sowohl mit geringem Aufwand neue EnKF-Methoden zu SHEMAT-Suite hinzuzufügen, als auch robuste Vergleiche zwischen implementierten EnKF-Methoden durchzuführen.

In der restlichen Arbeit nutze ich die vorgestellte Softwarestruktur zur Erforschung zweier Arten von Unsicherheiten, die bei einer Anwendung des EnKF auf Grundwasserströmungsberechnungen auftreten können. Zunächst wird die Zufälligkeit von Permeabilitätsschätzungen untersucht, die von der numerischen Zufallszahlenerzeugung herrührt. Insbesondere wird untersucht, wie sich diese Zufälligkeit auf Vergleiche von EnKF-Methoden und die Aussagekraft dieser Vergleiche auswirkt. Schließlich wird zur Unterdrückung unerwünschter, fälschlicherweise während einer Parameterschätzung entstehender Korrelationen eine neue EnKF-Methode namens Pilotpunkt-EnKF eingeführt.

SHEMAT-Suite In Kapitel 3 dieser Doktorarbeit wird die wissenschaftliche Software SHEMAT-Suite besprochen. SHEMAT-Suite ist ein offener Quellcode zur numerischen Simulation von Grundwasserströmung, Wärme- und Massentransport in porösen Medien. Darüberhinaus kann man SHEMAT-Suite zur Parameterschätzung und Datenassimilierung durch stochastische Ansätze wie Monte Carlo Methoden und den Ensemble Kalman Filter nutzen. In Kapitel 3 wird die wissenschaftliche Anwendung von SHEMAT-Suite anhand von Beispielen vorgestellt. Die folgenden Beispiele wissenschaftlicher Softwareentwicklung, die während der Arbeit an dieser Dissertation implementiert wurden, werden vorgestellt: die modulare Implementierung einer Suite von EnKF-Methoden und die Verwendung der Software git für das Aufteilen von SHEMAT-Suite in sogenannte Zweige, wobei jeder Zweig eine der Funktionalitäten von SHEMAT-Suite enthält. Diese Implementierungen wurden nach den folgenden wissenschaftlichen Kriterien durchgeführt: (1) Reproduzierbarkeit, (2) Falsifizierbarkeit, (3) Kommunikation der Ergebnisse und (4) eine tragfähige Grundlage für die künftige Arbeit. Das Kapitel wird abgeschlossen, indem Richtlinien für wissenschaftliche Softwareentwicklung vorgestellt werden.

Vergleich von EnKF-Methoden Kapitel 4 dieser Arbeit widmet sich der Untersuchung des Einflusses von Zufallszahlen auf die Permeabilitätsschätzung mit dem EnKF. Im Prinzip wird jede ensemble-basierte Parameterschätzung durch die Wahl von Zufallszahlen beeinflusst. Normalerweise kann die Ensemblegröße für eine Parameterschätzmethode so groß gewählt werden, dass der Einfluss der Zufallszahlen vernachlässigbar wird. Bei großskaligen Untergrundmodellen ist man jedoch in der Regel auf eine kleine Ensemblegröße beschränkt. In diesem Fall gewinnen die Schwankungen der Ergebnisse, die durch veränderte Zufallszahlen erklärt werden können, an Einfluss. In Kapitel 4 argumentiere ich, dass die Kenntnis des Zufallszahlen-Fehlers wesentlich für den Vergleich zwischen EnKF-Varianten ist. Um dies zu belegen, wird die folgende große Zahl an Vergleichen durchgeführt: Für sieben Ensemblegrößen (50, 70, 100, 250, 500, 1 000, 2 000) und sieben EnKF-Varianten (der gedämpfte, iterative, lokale, hybride, duale, Normal-Score und der klassische EnKF), werden 1 000 synthetische Parameterschätzungsexperimente für die folgenden zwei Anordnungen berechnet: ein 2-D Tracer-Transport und ein 2-D-Strömungsproblem mit einer Injektionsbohrung. Bisher werden in der Literatur selten mehrere synthetische Parameterschätzungsexperimente mit unterschiedlichen Zufallszahlen durchgeführt, wenn die Leistungsfähigkeit von EnKF-Methoden verglichen wird. Indem bei den synthetischen Parameterschätzungen nur die Zufallszahlen geändert werden, werden 1 000 geschätzte Permeabilitätsfelder je Ensemblegröße gesammelt. Diese geschätzten Permeabilitätsfelder werden mit einem synthetisch erzeugten wahren Permeabilitätsfeld verglichen, indem der Root-Mean-Square-Fehler berechnet wird (RMSE). Nach Vergleich der mittleren RMSEs für die sieben EnKF-Methoden bei der Ensemblegröße 50, komme ich zu dem Schluss, dass mindestens 10 synthetische Parameterschätzungsexperimente erforderlich sind, um RMSE-Differenzen zwischen EnKF-Varianten von kleiner als 10 % korrekt zu unterscheiden. Zur Unterscheidung von RMSE-Differenzen, die kleiner als 2 % sind, sind für die kleinen, aber üblichen Ensemblegrößen 50, 70, 100 und 250 bis zu 100 synthetische Parameterschätzungsexperimente erforderlich.

Pilotpunkt-Ensemble-Kalman-Filter In Kapitel 5 dieser Doktorarbeit wird der Pilotpunkt-Ensemble-Kalman-Filter (PP-EnKF) eingeführt. Der PP-EnKF ist eine Variante des Ensemble-Kalman-Filters, die ungewünschte Korrelationen unterdrückt. Ungewünschte oder fälschlich entstandene (englisch: spurious) Korrelationen beeinträchtigen die Parameterschätzung mit dem EnKF, wenn kleine Ensemblegrößen benutzt werden. Unter Verwendung des Vergleichs aus dem zweiten Teil der Arbeit, wird der PP-EnKF mit sieben anderen EnKF-Methoden verglichen. Für das Tracer-Transport-Modell steht der PP-EnKF an dritter Stelle der acht Methoden in einer gemittelten Gesamtwertung. Neben dem Vergleich der RMSEs, betrachte ich die Gesamtstandardabweichung. Die Gesamtstandardabweichung ist ein Maß dafür, ob die Verteilung der Permeabilitäten nach der Korrektur durch den EnKF breit genug bleibt. Für das 2-D-Strömungsproblem und für eine Ensemblegröße von 50 liefert der PP-EnKF die Gesamtstandardabweichung, die von allen EnKF-Methoden am nächsten an einer Referenz mit Ensemblegröße 10 000 liegt. Neben diesem Resultat liefert der PP-EnKF Korrelationsstrukturen, die deutlich näher an einer Referenz liegen als die des klassischen EnKF. Dies verdeutlicht die Fähigkeit des PP-EnKF zur Unterdrückung unerwünschter Korrelationen bei der Parameterschätzung unter Verwendung kleiner Ensemblegrößen. Aufgrund der guten Ergebnisse des PP-EnKF im Methodenvergleich schlage ich vor, den PP-EnKF künftig bei der Permeabilitätsschätzung in Grundwasserströmungssimulationen zu verwenden.

Schluss Die Ergebnisse dieser Arbeit wollen Hinweise zur effizienten Nutzung wachsender Computerressourcen auf dem Gebiet der Permeabilitätsabschätzung und der Grundwasserflussschätzung geben. In der Mehrzahl der wissenschaftlichen Disziplinen ist die Unsicherheitsquantifizierung von großer Bedeutung. Durch wachsende Computerleistung wird Unsicherheitsquantifizierung auch bei bisher zu aufwändigen Berechnungen möglich. Aus einer Reihe von Gründen ist die Unsicherheitsquantifizierung besonders wichtig zur Permeabilitätsabschätzung und bei Grundwasserströmungssimulationen: (1) Es gibt normalerweise viele ungenau bestimmte Größen in Untergrundmodellen, die einen Einfluss auf die Bestimmung der Grundwasserströmung ausüben, (2) Untergrundmodelle sind typischerweise groß und (3) folglich müssen kleine Ensemblegrößen für Sampling-Methoden zur Unsicherheitsquantifizierung ausreichen. In dieser Arbeit, wird eine Software vorgestellt, mit der die Sampling-Methode Ensemble Kalman Filter (EnKF) zur Parameterschätzung verwendet werden kann. Außerdem können mit der Software EnKF-Methoden verglichen werden. EnKF-Methoden werden im Rahmen der wissenschaftlichen Software SHEMAT-Suite so implementiert, dass die Implementierung wissenschaftliche Prinzipien erfüllen soll, z.B. Reproduzierbarkeit und Falsifizierbarkeit. Beim Vergleich von EnKF-Verfahren wird gezeigt, dass Zufallszahlen einen nicht zu vernachlässigenden Einfluss auf derartige Vergleiche haben. Schließlich wird der PP-EnKF eingeführt, eine EnKF-Variante zur Unterdrückung ungewünschter Korrelationen, die durch kleine Ensemblegrößen bei der Benutzung des EnKF entstehen.

If I have seen further it is by standing on the shoulders of Giants. But I make no question but you have divers very considerable experiments besides those you have published, it's very probable the same with some of those in my late papers.

Letter to Robert Hooke, 1675
Isaac Newton

1 Introduction

This introductory chapter provides background and literature on the simulation of groundwater flow and solute transport focusing on permeability estimation. In particular, literature on the ensemble Kalman filter (EnKF) and the software SHEMAT-Suite will be presented. For this thesis, several EnKF methods have been implemented in SHEMAT-Suite and this stochastic mode of SHEMAT-Suite is used for numerical simulation and data assimilation. A focus is laid on literature describing comparisons of various EnKF methods and the influence of spurious correlations on EnKF parameter estimation. In the second part of this chapter, the main objective of the thesis - presenting a framework for modular implementation and robust comparison of ensemble Kalman filters for parameter estimation in groundwater flow and transport - is presented and divided into sub-objectives that correspond to the subsequent chapters. Finally, a detailed outline of the thesis is given.

Contents

1.1 Background and literature

The dynamics of groundwater flow is important for many subsurface applications. For slow, laminar flow, Darcy's law and principles of mass continuity can be combined for deriving the groundwater flow equation and used for numerical modeling of groundwater flow (Bear, 1975, Marsily, 1986, Kinzelbach and Rausch, 1995). This thesis presents estimation techniques for the most important parameter of the groundwater flow equation: the permeability of the rock matrix. Before focusing on permeability estimation, I will list important applications of the groundwater flow equation. A classical application of the partial differential equation that describes groundwater

flow is reservoir engineering for oil recovery (Dake, 1983). Another application is the monitoring of aquifers, e.g. for water supply, groundwater remediation, or preventing contamination (Batu, 2010). The groundwater flow equation can be coupled with other partial differential equations, for example the heat equation (Carslaw and Jaeger, 1959) and the species transport equation (Zheng and Bennett, 2002, Clauser, 2003). These coupled partial differential equations can be solved numerically. As a large and sustainable energy source, geothermics is a useful application of the groundwater flow equation coupled with the heat equation. For economic as well as safety reasons, modeling groundwater and heat transport is a key step towards putting a geothermal system into operation (Clauser, 1997, Mottaghy and Dijkshoorn, 2012, Grant, 2013, Niederau et al., 2017, Michalski and Klitzsch, 2018, Hruška et al., 2019, Deb et al., 2019). Equally, simulating coupled groundwater flow and species transport is important for a number of applications, including modeling groundwater pollution (Bear and Verruijt, 1987), interpreting tracer tests (Vogt et al., 2012), modeling seawater intrusion in coastal aquifers (Thomas et al., 2019), and storing nuclear waste (Bear et al., 1993). All these applications that combine groundwater flow with other physical processes have in common that it is essential to be able to estimate the permeabilities in the model domain. In this thesis, permeability estimation is carried out and compared using numerical 2D simulations of groundwater flow, in some cases combined with species transport modeling.

The inverse estimation of permeabilities received a lot of attention in the groundwater literature of the last decades (e.g. Kitanidis and Vomvoris, 1983, Carrera and Neuman, 1986, RamaRao et al., 1995, Gómez-Hernández et al., 1997, Hendricks Franssen et al., 2003, Chen and Zhang, 2006, Kurtz et al., 2014). Since the 1980s, a strong focus has been laid on the use of stochastic methods for solving inverse problems. Stochastic methods allow to quantify the uncertainty associated with the inversely estimated parameters. Non-linear models of the subsurface with many unknown parameters, as typical for groundwater simulations, pose challenges to stochastic inverse modeling. Sequential methods like the ensemble Kalman filter (EnKF, Evensen, 2003) were found to be relatively efficient for highly discretized models, still allowing for robust uncertainty quantification in spite of many unknowns. For this reason, there is a growing number of studies on the EnKF in groundwater hydrology (Chen and Zhang, 2006, Hendricks Franssen and Kinzelbach, 2008, Nowak, 2009, Vogt et al., 2012). In this thesis, variants of the EnKF are compared and a novel EnKF method is introduced.

Stochastic inversion methods are derived from probability theory. For instance, the usage of the EnKF for parameter estimation can be interpreted as a simplified Bayesian inference (Chen, 2003, Evensen, 2009). The simplification of the EnKF update step consists in the assumption of Gaussianity for the probability distributions that are part of the update step of the EnKF. Before discussing applications of the EnKF for stochastic data assimilation and parameter estimation, it is important to quote some literature regarding the mathematical background of the methods used in this thesis. For general probability theory, Loève (1977) gives a rigorous mathematical introduction, and Papoulis and Pillai (2002) give a modern overview of the basics. The foundational work for probability theory in the Bayesian framework was published by Jaynes (2003). A Bayesian view on data assimilation can be found in Sivia and Skilling (2006). Finally,

the literature on inverse problems is of interest for the reader of this thesis. Aster et al. (2013) give an introduction to general inverse problems. Tarantola (2005) uses Bayesian reasoning and provides mathematical background. Finally, Kaipio and Somersalo (2006) define best practices for computational inverse problems.

The EnKF, the stochastic method used in this thesis, is an ensemble-based version of Kalman filtering. Kalman filtering was originally introduced in the 1960s as a computationally advantageous state-space method for optimal estimation (Kalman, 1960, Jazwinski, 1970, Gelb, 1974). For large-scale meteorological applications, the ensemble Kalman filter was developed in the mid-90s (Evensen, 1994, Burgers et al., 1998). In the atmospheric sciences, the EnKF is mostly used as a filter for updating the initial states of dynamic variables (Houtekamer and Mitchell, 1998, Anderson, 2001, Hamill and Snyder, 2000, Kalnay, 2002). In the early 2000s, the EnKF was introduced to land surface and subsurface applications, including land surface modeling (Reichle et al., 2002, De Lannoy and Reichle, 2016), reservoir engineering (Lorentzen et al., 2003, Naevdal et al., 2005, Aanonsen et al., 2009, Vogt et al., 2012), as well as surface and subsurface hydrology (e.g. Vrugt et al., 2005, Chen and Zhang, 2006, Devegowda et al., 2010, Shi et al., 2014, Baatz et al., 2017). For land surface and subsurface applications, model parameters are often the main source of uncertainty. In such cases, it is useful to apply the EnKF as a mixture of dynamic state optimization and parameter estimation (Naevdal et al., 2005, Hendricks Franssen and Kinzelbach, 2008). In this thesis, the EnKF is used for subsurface permeability estimation. In particular, I address possible obstacles when using the EnKF and I quantify the influence of these obstacles for permeability estimation.

The first step for uncertainty quantification with the EnKF is its implementation and in particular its integration into numerical software. The groundwater flow, heat and solute transport simulations of this thesis are computed using SHEMAT-Suite, a numerical software for simulating flow, heat and species transport in porous media. In Chapter 3, the software SHEMAT-Suite is presented, alongside a description of the changes introduced to the software during the work on this thesis and the concepts of scientific programming behind these changes. Other software packages can also handle tasks similar to SHEMAT-Suite. Some are available as purchasable research code or as commercial packages, such as TOUGH (Finsterle et al., 2014), FEFLOW (Diersch, 2014) and STOMP (White et al., 2008), others are distributed as open-source initiatives, including OpenGeoSys (Kolditz et al., 2012), DuMuX (Flemisch et al., 2011) and MODFLOW (Hughes et al., 2017). SHEMAT-Suite complements available open-source packages by a powerful toolkit of functionalities including: (a) multi-level OpenMP parallelization for simulations of flow through porous media (Wolf et al., 2008, Bücker et al., 2009, Wolf, 2011); (b) a modular user-framework for implementing subsurface parameter dependencies on pressure and temperature, and customized graphical output; (c) species transport including density-driven flow; (d) uncertainty quantification, parameter estimation and data assimilation by stochastic approaches (Monte Carlo, ensemble Kalman filter as shown in this thesis) and also by deterministic Bayesian inversion based on Automatic Differentiation (AD) for calculating exact Jacobian matrices of the forward model (Rath et al., 2006).

SHEMAT-Suite in its present form has evolved from SHEMAT (Simulator for HEat and MAss

Transport, Clauser, 1988, 2003), a fully coupled flow and heat transport model that was developed from the isothermal USGS 3-D groundwater model of Trescott and Larson (1976, 1977). Originating from the Fortran-77 code SHEMAT, SHEMAT-Suite has been rewritten in Fortran-95 (Rath et al., 2006), refactored and enhanced subsequently with various additional capabilities within numerous research and PhD projects. These capabilities include (1) a parallelization scheme for SHEMAT-Suite (Wolf, 2011); (2) the solution of the systems of coupled partial differential equations by implicit Picard iteration in a finite difference discretization (Lynch, 2005, Huyakorn, 2012); (3) a direct linear solver for small simulations, an implementation of the Bi-Conjugate Gradients Stabilized (Bi-CGStab) method for large simulations (van der Vorst, 1992); (4) additional solvers and preconditioners easy to integrate in SHEMAT-Suite owing to its modular structure of the solver implementation; (5) linear algebra functions used from the libraries BLAS (Lawson et al., 1979, Dongarra et al., 1988, 1990) and LAPACK (Anderson et al., 1999); (6) output of variable and parameter fields possible in many formats, most notably VTK (Schroeder et al., 2006) and HDF5[1] (The HDF Group, 1997-2022); (7) deterministic Bayesian inversion (Rath et al., 2006); (8) automatic differentiation of the forward code by the software Tapenade (Hascoët and Pascual, 2013) for computing the Jacobian in the deterministic inversion framework; (9) sequential Gaussian simulation (SGSim[2]) from the geostatistical library GSLib (Deutsch and Journel, 1995) for random field generation in the stochastic simulation mode within a Monte Carlo framework; (10) Bayesian updates implemented as ensemble Kalman filter (EnKF[3]) analysis equations (Evensen, 2003). The current, open-source version of SHEMAT-Suite is restricted to a single fluid phase in the porous medium. However, a multi-phase fluid version, which can, e.g., simulate CO_2, steam (Büsing et al., 2014, Büsing, 2020), and supercritical water (Büsing et al., 2016), is under development.

A versatile and openly available numerical simulation package such as SHEMAT-Suite is of significance for fostering research in the field of geothermal energy. Especially in the context of climate change, geothermal energy is becoming increasingly important as a renewable energy source worldwide. Other important application areas of SHEMAT-Suite are groundwater management and contaminant transport simulations. These application areas of modeling groundwater flow, heat and species transport are reflected by the scientific contributions of SHEMAT-Suite so far, which can be divided into four categories: geothermics, paleoclimate, hydrogeology and, additionally, inverse method development. Geothermics is the most important application of SHEMAT-Suite and motivated the software development originally. The applications of SHEMAT-Suite in geothermics range from large-scale geothermal simulations (Vogt et al., 2012, Ebigbo et al., 2016) to small-scale borehole heat exchanger and temperature sensor simulations (Michalski and Klitzsch, 2018). The second field of applications concerns paleoclimate related studies, including an ice module used for simulating freezing and thawing processes in porous media (e.g. Mottaghy and Rath, 2006). Additionally, SHEMAT-Suite has been used in studies of broad geological interest, including a simulation of submarine groundwater discharge at the New Jersey shelf, USA, since the last ice age (Thomas et al., 2019) and a simulation of convection

[1] https://www.hdfgroup.org/downloads/hdf5/ [Accessed 2020/02/19]
[2] http://www.gslib.com/gslib_help/sgsim.html [Accessed 2020/02/26]
[3] https://github.com/geirev/EnKF_analysis [Accessed 2020/02/19]

cells in the Perth Basin, Australia (Niederau et al., 2017). General development of algorithms has been carried out using SHEMAT-Suite. For example, SHEMAT-Suite uses different techniques from high-performance computing (Rostami and Bücker, 2014, Rostami et al., 2014) and is coupled to the optimal experimental design framework EFCOSS (Rasch and Bücker, 2010) to enhance reservoir property estimation (Seidler et al., 2016). The work on this thesis began with the software SHEMAT-Suite providing the possibility of data assimilation with the EnKF, potentially a damped EnKF. Building on this code, the software was restructured in a modular fashion and the suite of EnKF methods was implemented that constitutes the SHEMAT-Suite Stochastic Mode and that is used in Chapter 4 and 5 of this thesis. In Chapter 3 of this thesis, SHEMAT-Suite as a whole is presented, putting a strong focus on the Stochastic Mode and trying to extract some general guidelines for scientific software development that can be derived from the work on this thesis.

Performing meaningful comparisons of EnKF methods is one obstacle for uncertainty quantification using the EnKF. As the number of EnKF methods is growing, performance comparisons between these methods gain importance as well. In this thesis, I argue that many performance comparisons are insignificant, meaning that the difference in statistical performance measures like the root-mean-square error (RMSE) between tested methods is smaller than the uncertainty of these performance measures. Performance comparisons of EnKF methods can be based on so-called synthetic experiments. I define a single synthetic experiment as a comparison of a number of data assimilation methods for a given groundwater flow problem and for a given set of synthetically generated observations. Thus, a single synthetic experiment consists of multiple data assimilation experiments, where the number of data assimilation experiments is equal to the number of data assimilation methods that are compared. For such a synthetic experiment, it is important that the complete set of inputs is identical for each data assimilation method. The set of inputs includes the statistical random seeds that I focus on in this thesis. Random seeds are used for sampling probability distributions during a synthetic experiment. Quantities with sampled probability distributions in a synthetic experiment include initial random variables and parameters, as well as observations. For large ensemble sizes, the influence of the random seed vanishes. However, varying the random seeds may lead to significant variations of the outcome of a synthetic experiment for the smaller ensemble sizes that are typical in groundwater flow simulations. Another example for an input of a synthetic experiment is the ensemble of initial permeability fields, in particular the size of the ensemble and the single realizations.

For a performance comparison of the data assimilation methods, it is beneficial to use multiple synthetic experiments. From such a set of synthetic experiments, the uncertainty of statistical performance measures can be derived, when only the statistical random seeds are varied across synthetic experiments, while all other inputs remain constant. There are several examples of EnKF comparison studies in the hydrologic sciences and they used varying numbers of synthetic experiments. When performance measures such as the root-mean-square error (RMSE) or the mean absolute error are considered, differences between methods often are smaller than 10 %. For examples of performance comparisons between EnKF methods, for which at least a part of the differences between performance measures are smaller than 10 %, see, e.g., Moradkhani

et al. (2005), Camporese et al. (2009), Sun et al. (2009b), Hendricks Franssen and Kinzelbach (2009), Zhou et al. (2011), El Gharamti and Hoteit (2014), El Gharamti et al. (2014, 2015), Liu et al. (2016). The comparison of EnKF algorithms is often based on less than 10 synthetic experiments. In this thesis, I will show that this is in general not enough. Moreover, when employing the definition of a synthetic experiment given above, most studies are based on single synthetic experiments. Some studies already evaluate a large number of synthetic experiments for performance comparisons. Moradkhani et al. (2005) employed 500 synthetic experiments of ensemble size 50, Sun et al. (2009a) used 16 synthetic experiments and Schöniger et al. (2012) computed 200 test cases. In Chapter 4 of this thesis, two simple subsurface models are considered for systematically evaluating the number of synthetic experiments needed to distinguish performances of EnKF variants.

A second obstacle for uncertainty quantification using the EnKF are so-called spurious correlations. In the update step of the EnKF, correlations between all variables and parameters are considered. The correlations that are used for the EnKF are ensemble-based, i.e. computed from an ensemble of variables and parameters. Therefore, these correlations may suffer from sampling error. Correlations that wrongly influence the EnKF update due to sampling errors are called spurious. It was observed in several studies that small ensemble sizes lead to spurious correlations among other problems. Spurious correlations can result in filter inbreeding (Houtekamer and Zhang, 2016, Zovi et al., 2017) and ultimately filter divergence (Hamill et al., 2001). Filter inbreeding occurs when the spread of the updated ensemble becomes so small that it does not properly represent the error of the EnKF updates. Filter divergence means that the updates of the ensemble center around a wrong parameter value, even though the spread of the ensemble minimizes as if the filter would be approaching the correct parameter value. Filter divergence can result from filter inbreeding. Many methods have been developed to tackle spurious correlations. In Chapter 4 of this thesis, I compare the following methods: damped EnKF (Hendricks Franssen and Kinzelbach, 2008), local EnKF (Hamill et al., 2001), hybrid EnKF (Hamill and Snyder, 2000), dual EnKF (Moradkhani et al., 2005, El Gharamti et al., 2013), iterative EnKF (Sakov et al., 2012) and normal score EnKF (Zhou et al., 2011, Schöniger et al., 2012, Li et al., 2012). Two examples of modern localization approaches aimed at reducing spurious correlations are presented by Chen and Oliver (2010) and Emerick and Reynolds (2011). Methods using decompositions of probability distributions into sums of Gaussian distributions have been introduced by Sun et al. (2009b) and Liu et al. (2016). The last two methods are not evaluated in this thesis due to limited resources. In Chapter 5 of this thesis, I propose the pilot point ensemble Kalman filter, a method tailored to reducing spurious correlations.

1.2 Objectives and outline

The overall objective of this thesis is presenting a framework for modular implementation and robust comparison of ensemble Kalman filters for parameter estimation in groundwater flow and species transport. Three objectives are derived from this overall objective, namely implementing EnKF methods, comparing EnKF methods, and introducing a new EnKF method:

- Objective 1 (Chapter 3): Implementing a suite of EnKF methods and integrating the suite of methods into a numerical software for groundwater flow and species transport, thereby proposing guidelines for scientific programming, such as reproducibility and falsifiability.
- Objective 2 (Chapter 4): Comparing performance of EnKF methods and demonstrating the importance of considering the uncertainty of performance measures related to a limited number of synthetic experiments.
- Objective 3 (Chapter 5): Introducing the pilot point ensemble Kalman filter as a fast and robust EnKF method for parameter estimation. The pilot point ensemble Kalman filter (PP-EnKF) reduces spurious correlations and partly prevents an underestimation of posterior variances of parameters.

Chapter 2 contains the general theory that is a prerequisite for understanding the methods in the subsequent chapters of this thesis. This general theory is divided into two parts. The first part of the theory chapter presents the model equations of SHEMAT-Suite, the scientific software used for the simulations presented in this thesis. Additionally, the discretization of the groundwater flow equation is presented as it is implemented in SHEMAT-Suite. The second part of the theory chapter introduces the Kalman filter and its efficient Monte Carlo variant, the ensemble Kalman filter (EnKF).

Chapter 3 treats the implementation of EnKF methods in a refactored version of the scientific software SHEMAT-Suite. SHEMAT-Suite is introduced as an open-source numerical code for simulating flow, heat and species transport in porous media. After motivating the significance of SHEMAT-Suite, the software is described in detail, equations and discretization are explained and illustrative examples are given. Then, the focus is shifted to the implementation of the EnKF methods in SHEMAT-Suite that has been carried out during the work on this thesis and that are used in Chapters 4 and 5. Finally, some guidelines for using the version control software git for software development are derived from the development of SHEMAT-Suite that was carried out during this thesis. These guidelines promote scientific principles, including reproducibility and falsifiability.

Chapter 4 treats the influence of the sampling error on permeability estimation in EnKF studies. An extensive performance comparison of EnKF methods is carried out. For seven ensemble sizes (50, 70, 100, 250, 500, 1 000, 2 000) and seven EnKF variants (damped, iterative, local, hybrid, dual, normal score, and classical EnKF), 1 000 synthetic parameter estimation experiments are computed for two setups: a 2-D tracer transport problem and a 2-D flow problem with one injection well. Using this large number of synthetic experiments, I estimate how many synthetic experiments are needed for a robust comparison of different EnKF variants. The main performance measure that is evaluated is the root-mean-square error (RMSE). Results include the mean values of the RMSE for the full number of synthetic experiments. The question of the robustness of performance comparisons is subsequently tackled by examining smaller subsets of synthetic experiments. Finally, I check certain influences on the performance comparisons, including ensemble size, specific parameter settings in the EnKF methods, and observation noise.

Chapter 5 treats spurious correlations, a sampling error known to influence EnKF parameter estimation. In this chapter, the pilot point ensemble Kalman filter (PP-EnKF) is derived. The PP-EnKF is a variant of the ensemble Kalman filter adapted to suppressing spurious correlations. The PP-EnKF is compared to the hybrid EnKF and the local EnKF, two methods that were introduced in the past for suppressing spurious correlations. For this comparison, the comparison setup from Chapter 4 is reused. The PP-EnKF has two main inputs: the pilot point grid and the prior correlation structure of the parameters. The pilot point grid specifies the subset of parameters that are subject to the EnKF update before a kriging interpolation. The prior parameter correlation structure is an input of the kriging interpolation. Due to the importance of these inputs, the comparison of the PP-EnKF to other EnKF methods was extended to test for the role of the prior estimate of the correlation length and the design of the pilot point grid. Another objective is to check the ability of the PP-EnKF to prevent an underestimation of the posterior variance. To this end, overall standard deviations of the updated permeability fields are compared. Additionally, an analysis of the correlations between observations and permeabilities throughout the model domain is presented. Comparing these correlations from synthetic experiments with small ensemble sizes to reference correlations from a synthetic experiment using a large ensemble size, I highlight the reduction of spurious correlations by the PP-EnKF compared to the classical EnKF.

Chapter 6 presents general conclusions focusing on the main objectives of this thesis. This is followed by more detailed conclusions relating to the sub-objectives of Chapters 3 to 5. The thesis is concluded by giving an outlook on uncertainty quantification in subsurface modeling and showing how the work presented in this thesis may support future developments in this field.

Et harum scientiarum porta et clavis est mathematica, quam sancti a principio mundi invenerunt, ut ostendam, et quae semper fuit in usu omnium sanctorum et sapientum prae omnibus aliis scientiis.

Opus Majus,
Pars Quarta, Distincto Prima, Capitulum I
Roger Bacon

2

Theory and methods

In this chapter, mathematical background for the two main topics underlying the research of this thesis is presented: simulation of flow, heat and tracer transport using SHEMAT-Suite and parameter estimation using the ensemble Kalman filter (EnKF).

Contents

2.1 Flow, heat and tracer simulation using SHEMAT-Suite

In this section, the governing equations of SHEMAT-Suite are presented. For mathematical background on partial differential equations, see Arnold (2004), Evans (2010), Courant and Hilbert (1924), and in particular on a mathematical treatment of fluid flow equations see Christie and Aris (1965) and Chorin and Marsden (1979). For groundwater flow and flow in porous media see Bear (1975), Marsily (1986), and Kinzelbach and Rausch (1995). For information on the implicit solution of the partial differential equations implemented in SHEMAT-Suite, see Clauser (2003), Lynch (2005), Huyakorn (2012), and Rath et al. (2006).

2.1.1 Governing equations

Simulation of laminar groundwater flow is based on Darcy's law for the computation of the specific discharge $\mathbf{v}$ $[LT^{-1}]$:

$$\mathbf{v} = -\frac{\rho_f g}{\mu_f}\mathbf{K} \cdot \left(\vec{\nabla} h + \rho_r \vec{\nabla} z\right). \tag{2.1}$$

Here g $[LT^{-2}]$ is the gravitational constant, ρ_f $[ML^{-3}]$ denotes fluid density, μ_f $[ML^{-1}T^{-1}]$ is the dynamic viscosity of the fluid, and $\mathbf{K}$ $[L^2]$ is the hydraulic permeability tensor. The permeability $\mathbf{K}$ is a property of the porous medium. h $[L]$ is the reference hydraulic head. ρ_r $[-]$ is the relative fluid density variation with respect to a reference density ρ_0 $[ML^{-3}]$:

$$\rho_r = \frac{\rho_f - \rho_0}{\rho_0}. \tag{2.2}$$

Finally, z $[L]$ is the elevation above a reference datum. This reference datum may be chosen as the sea level in groundwater models. For computational reasons, the reference datum can also be chosen as the deepest depth of the model domain (Clauser, 2003, Bear, 1975). In Equation (2.1), z is a function of the location in the model domain.

Given the reference fluid density ρ_0, the reference hydraulic head h is computed using as follows:

$$h = z + \frac{P}{\rho_0 g}. \tag{2.3}$$

Here, P $[ML^{-1}T^{-2}]$ is the pressure at the same location, where the reference hydraulic head h and the height z are evaluated. Conceptually, the reference hydraulic head describes the height to which water in a hypothetical well rises due to the fluid pressure P given a constant reference fluid density ρ_0 for the hypothetical well. This assumption simplifies the conversion between reference hydraulic head and pressure. The observed hydraulic head can be computed from the reference hydraulic head by considering the fluid density variations. For the sake of brevity, the word *reference* will from now on be dropped and the reference hydraulic head will be called hydraulic head unless it is important to stress that the hydraulic head is a reference hydraulic head.

To summarize the various influences on the specific discharge $\mathbf{v}$, permeability characterizes the influence of the rock matrix on hydraulic conductivity, while fluid density ρ_f and the fluid dynamic viscosity μ_f characterize the influence of the fluid. The hydraulic conductivity tensor transforms the gradient of the hydraulic head and the buoyancy term into the specific discharge.

The partial differential equation that is at the center of this thesis is the transient groundwater flow equation for laminar flow in a porous medium. The groundwater flow equation describes the dynamics of the reference hydraulic head h under the influence of varying specific discharges and sinks or sources:

$$S_s \frac{\partial h}{\partial t} = -\nabla \cdot \mathbf{v} + W. \tag{2.4}$$

Here, t $[T]$ denotes time, and S_s $[L^{-1}]$ is the specific storage, a function of the fluid density ρ_f $[ML^{-3}]$ and of the porosity φ $[-]$ of the rock matrix (see Clauser, 2003).

The dynamics of the hydraulic head are determined by the divergence of the specific discharge $\mathbf{v}$ and the source term W $[T^{-1}]$. The source term W is specified as an external volume flow W_V $[L^3T^{-1}]$ that is normalized by a control volume dV $[L^3]$. In computations from this work, the control volume is the size of the discretization cell that is subject to the external volume flow. To obtain the source term W, the normalized volume flow is multiplied by a quotient of the source fluid density ρ^* $[ML^{-3}]$ and the reference density ρ_0 $[ML^{-3}]$:

$$W = \frac{W_V}{dV}\frac{\rho^*}{\rho_0}, \tag{2.5}$$

where the density quotient ensures that the source term enters the groundwater flow equation in terms of the reference hydraulic head (Clauser, 2003).

Now, some modeling choices concerning the groundwater flow equation valid throughout this thesis are highlighted. The fluid will be water. The water density ρ_f, the gravitational constant g, and the dynamic viscosity of water μ_f are assumed as constant throughout the dimensions of an aquifer model. The hydraulic permeability tensor $\mathbf{K}$ is assumed to be isotropic. In the isotropic case, a scalar permeability specifies the full tensor. In this thesis, the logarithmic form of the scalar permeability $K_{\log}$ $[-]$ is mostly used. The logarithmic scalar permeability is obtained by first using the scalar value of $\mathbf{K}$ in the unit m^2 and then taking the logarithm of ten. For scalar permeability, the direction of $\mathbf{v}$ is fully determined by the two gradients of Equation (2.1). In the horizontal case, the direction of $\mathbf{v}$ corresponds to the direction of the steepest descent of the hydraulic head field.

In absence of a source term, the equation for species transport is given by the advection-dispersion equation:

$$\varphi\frac{\partial c}{\partial t} = \nabla\left(\mathbf{D}\cdot\nabla c\right) - \nabla\cdot\left(c\mathbf{v}\right), \tag{2.6}$$

where c $[ML^{-3}]$ is the concentration of a tracer. The time derivative of the concentration is determined by two terms, a dispersion-diffusion term and an advection term. The diffusion in the concentration field itself is influenced by the hydrodynamic dispersion tensor $\mathbf{D}$ $[L^2T^{-1}]$. The advection is driven by the specific discharge $\mathbf{v}$. The specific discharge $\mathbf{v}$ couples the species transport equation to the groundwater flow equation. The porosity of the rock matrix φ $[-]$ is assumed as constant. In models from this work the water flow is so strong, that the hydrodynamic dispersion tensor $\mathbf{D}$ has a smaller influence on concentration evolution than the specific discharge from the Darcy equation.

For completeness, the heat transport equation is now shortly discussed, again without sources. The heat transport equation is central to SHEMAT-Suite as a software applied for geothermal energy exploration. In this thesis, the heat transport equation is not used for permeability estimation. However, it should be noted that there are efforts in this respect by using temperature measurements for estimating groundwater flow (Michalski and Klitzsch, 2018, Seidler et al., 2016).

The heat transport equation is given by:

$$(\varphi\rho_f c_f + (1-\varphi)\rho_m c_m)\,\frac{\partial T}{\partial t} = \nabla\,(\lambda \cdot \nabla T) - \nabla \cdot (\rho_f c_f T \mathbf{v})\,. \tag{2.7}$$

Like the species transport equation, this is an advection diffusion equation. The dynamics of the temperature T $[\Theta]$ are determined by the diffusion and the advection summands. Diffusion is influenced by the tensor of thermal conductivities of the matrix-fluid mixture λ $[MLT^{-3}\Theta^{-1}]$. Advection is influenced by the fluid density ρ_f and the fluid specific heat capacity c_f $[L^2T^{-2}\Theta^{-1}]$. Finally, there are two rock properties in the equation, the matrix density ρ_m $[ML^{-3}]$ and the matrix specific heat capacity c_m $[L^2T^{-2}\Theta^{-1}]$. In this thesis, SHEMAT-Suite is used for simulating transient groundwater flow and species transport. The temperature is assumed to be constant during these simulations and I do not consider heat sources or sinks.

2.1.2 Discretization of the groundwater flow equation

The software SHEMAT-Suite (Rath et al., 2006) uses the finite difference method for discretization in order to solve the governing equations for groundwater flow, heat and species transport. The discretization of SHEMAT-Suite that will be explained in some detail in this section, is largely the same as for its predecessor SHEMAT (Clauser, 2003) and based for example on Zheng and Bennett (2002). As in SHEMAT, the governing equations are discretized using the Crank-Nicolson method (Crank and Nicolson, 1947). In the time domain, this means using backward differences and thus the equations are solved implicitly. In the space domain, the discretization is carried out according to a block-centered finite difference scheme. Time and space discretization result in a system of linear equations that is solved numerically. Here, the fundamentals that are presented in Lynch (2005) and Huyakorn (2012) are not recalled. Instead, I will focus on the following: a thorough discussion of the discretization of the groundwater flow equation exactly as it is implemented in the current version of SHEMAT-Suite. This discussion will be carried out in the following steps. First, the groundwater flow equation is discretized. Secondly, all parameters in the discretization are traced back to input parameters of SHEMAT-Suite. Finally, the implementation of the discretization in SHEMAT-Suite is described.

The groundwater flow equation in the form that is discretized in SHEMAT-Suite is obtained by inserting the specific discharge from Equation (2.1) in the groundwater flow equation (Equation (2.4)):

$$S_s\frac{\partial h}{\partial t} = \nabla \cdot \left(\frac{\rho_f g}{\mu_f}\mathbf{K} \cdot \left(\vec{\nabla}h + \rho_r\vec{\nabla}z\right)\right) + W. \tag{2.8}$$

For simplicity, the groundwater flow equation is expressed in terms of the hydraulic conductivity $\mathbf{k}$ $[LT^{-1}]$:

$$S_s\frac{\partial h}{\partial t} = \nabla \cdot \left(\mathbf{k} \cdot \left(\vec{\nabla}h + \rho_r\vec{\nabla}z\right)\right) + W. \tag{2.9}$$

In the following, the discretization of the groundwater flow equation will be explained in two steps (see Figure 2.1 for the relation of important discretized quantities to the grid). The following paragraphs are an in-depth explanation and elaboration of the discretization of the groundwater

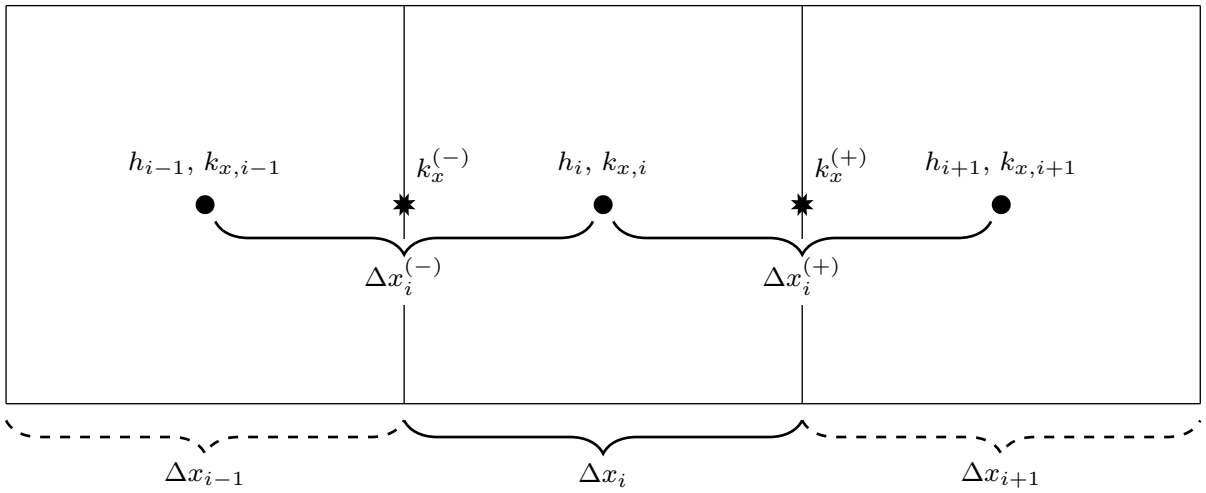

Figure 2.1: Exemplary grid locations of local values for hydraulic head h and hydraulic conductivity k_x. Additionally, the x-increments of the discretization of the groundwater flow equation in SHEMAT-Suite are shown. The indices j and k of the y- and z-direction have been dropped for simplicity.

flow equation that can be found in condensed form in Clauser (2003). There, one can also find a lot of additional information on the numerical techniques of SHEMAT and SHEMAT-Suite. All derivatives in the groundwater flow equation are discretized at a general cell with directional indices i, j, k. In the first step of the discretization, the time derivative and the outer spatial derivative, i.e. the divergence, are discretized by introducing difference quotients.

$$
\begin{aligned}
S_s \frac{h_{i,j,k} - \hat{h}_{i,j,k}}{\Delta t} - W = & \\
& \frac{1}{\Delta x_i} \left[\left(\mathbf{k} \cdot \left(\vec{\nabla} h + \rho_r \vec{\nabla} z \right) \right)_x^{(+)} - \left(\mathbf{k} \cdot \left(\vec{\nabla} h + \rho_r \vec{\nabla} z \right) \right)_x^{(-)} \right] \\
& + \frac{1}{\Delta y_j} \left[\left(\mathbf{k} \cdot \left(\vec{\nabla} h + \rho_r \vec{\nabla} z \right) \right)_y^{(+)} - \left(\mathbf{k} \cdot \left(\vec{\nabla} h + \rho_r \vec{\nabla} z \right) \right)_y^{(-)} \right] \\
& + \frac{1}{\Delta z_k} \left[\left(\mathbf{k} \cdot \left(\vec{\nabla} h + \rho_r \vec{\nabla} z \right) \right)_z^{(+)} - \left(\mathbf{k} \cdot \left(\vec{\nabla} h + \rho_r \vec{\nabla} z \right) \right)_z^{(-)} \right] \\
& = \frac{1}{\Delta x_i} \left[k_x^{(+)} \left(\frac{\partial h_i}{\partial x} \right)^{(+)} - k_x^{(-)} \left(\frac{\partial h_i}{\partial x} \right)^{(-)} \right] \\
& \quad + \frac{1}{\Delta y_j} \left[k_y^{(+)} \left(\frac{\partial h_j}{\partial y} \right)^{(+)} - k_y^{(-)} \left(\frac{\partial h_j}{\partial y} \right)^{(-)} \right] \\
& \quad + \frac{1}{\Delta z_k} \left[k_z^{(+)} \left(\frac{\partial h_k}{\partial z} + \rho_r \right)^{(+)} - k_z^{(-)} \left(\frac{\partial h_k}{\partial z} + \rho_r \right)^{(-)} \right]
\end{aligned}
$$

Five things may be noted at this point. First, $\hat{h}_{i,j,k}$ denotes the known hydraulic head at cell (i, j, k) at the simulation time of the previous time step, with a time difference of Δt [T] to the current simulation time. Secondly, only the directional indices that change in a certain direction are specified in the corresponding summand (i in x-direction, j in y-direction, k in z-

direction). Thus, for example h_i is equivalent to $h_{i,j,k}$, and h_{i+1} is equivalent to $h_{i+1,j,k}$. Thirdly, the quantities with a (+) are located on the boundary of a grid cell in positive x/y/z-direction, while the quantities with a (−) are located on the boundary of a grid cell in negative x/y/z-direction (compare Figure 2.1). Fourth, as suggested by the (+)/(−)-notation, the endpoints of the difference quotients for the divergence are chosen at the cell boundaries. Finally, I assume that the hydraulic conductivity $\mathbf{k}$ is diagonal with diagonal elements k_x, k_y, k_z. This assumption is required by the open-source SHEMAT-Suite. However, relaxing the assumption of diagonal permeability tensors is part of the closed-source research surrounding SHEMAT-Suite (Chen et al., 2016).

In the second step, the inner partial derivatives, the gradients, are discretized. Contrary to the difference quotient used for the outer partial derivative, the endpoints of the difference quotient for the gradient are now located at the grid cell centers. This is in agreement with the discretization of second partial derivatives as shown in Huyakorn (2012).

$$\begin{aligned} S_s \frac{h_{i,j,k} - \hat{h}_{i,j,k}}{\Delta t} - W = & \frac{1}{\Delta x_i} \left[k_x^{(+)} \left(\frac{h_{i+1} - h_i}{\Delta x_i^{(+)}} \right) - k_x^{(-)} \left(\frac{h_i - h_{i-1}}{\Delta x_i^{(+)}} \right) \right] \\ & + \frac{1}{\Delta y_j} \left[k_y^{(+)} \left(\frac{h_{j+1} - h_j}{\Delta y_j^{(+)}} \right) - k_y^{(-)} \left(\frac{h_j - h_{j-1}}{\Delta y_j^{(-)}} \right) \right] \\ & + \frac{1}{\Delta z_k} \left[k_z^{(+)} \left(\frac{h_{k+1} - h_k}{\Delta z_k^{(+)}} + \rho_r^{(+)} \right) \right. \\ & \left. - k_z^{(-)} \left(\frac{h_k - h_{k-1}}{\Delta z_k^{(-)}} + \rho_r^{(-)} \right) \right] \end{aligned} \tag{2.10}$$

The hydraulic conductivities $k_{x/y/z}^{(+/-)}$ are computed at the cell boundaries. Inputs for computing the hydraulic conductivities are specified at the grid cell centers in SHEMAT-Suite. The hydraulic conductivities at the grid cell boundaries are derived from the hydraulic conductivities at the grid cell centers using the following harmonic averages:

$$k_x^{(+)} = \left[\frac{\left(\frac{\Delta x_{i+1}/2}{\Delta x_i^{(+)}} \right)}{k_{x,i+1}} + \frac{\left(\frac{\Delta x_i/2}{\Delta x_i^{(+)}} \right)}{k_{x,i}} \right]^{-1} \qquad k_x^{(-)} = \left[\frac{\left(\frac{\Delta x_i/2}{\Delta x_i^{(-)}} \right)}{k_{x,i}} + \frac{\left(\frac{\Delta x_{i-1}/2}{\Delta x_i^{(-)}} \right)}{k_{x,i-1}} \right]^{-1}$$

$$k_y^{(+)} = \left[\frac{\left(\frac{\Delta y_{j+1}/2}{\Delta y_j^{(+)}} \right)}{k_{y,j+1}} + \frac{\left(\frac{\Delta y_j/2}{\Delta y_j^{(+)}} \right)}{k_{y,j}} \right]^{-1} \qquad k_y^{(-)} = \left[\frac{\left(\frac{\Delta y_j/2}{\Delta y_j^{(-)}} \right)}{k_{y,j}} + \frac{\left(\frac{\Delta y_{j-1}/2}{\Delta y_j^{(-)}} \right)}{k_{y,j-1}} \right]^{-1}$$

$$k_z^{(+)} = \left[\frac{\left(\frac{\Delta z_{k+1}/2}{\Delta z_k^{(+)}} \right)}{k_{z,k+1}} + \frac{\left(\frac{\Delta z_k/2}{\Delta z_k^{(+)}} \right)}{k_{z,k}} \right]^{-1} \qquad k_z^{(-)} = \left[\frac{\left(\frac{\Delta z_k/2}{\Delta z_k^{(-)}} \right)}{k_{z,k}} + \frac{\left(\frac{\Delta z_{k-1}/2}{\Delta z_k^{(-)}} \right)}{k_{z,k-1}} \right]^{-1},$$

where the distances between neighboring grid cell centers given by

$$\Delta x_i^{(+)} = \frac{1}{2}\left(\Delta x_{i+1} + \Delta x_i\right) \qquad \Delta x_i^{(-)} = \frac{1}{2}\left(\Delta x_i + \Delta x_{i-1}\right)$$
$$\Delta y_j^{(+)} = \frac{1}{2}\left(\Delta y_{j+1} + \Delta y_j\right) \qquad \Delta y_j^{(-)} = \frac{1}{2}\left(\Delta y_j + \Delta y_{j-1}\right)$$
$$\Delta z_k^{(+)} = \frac{1}{2}\left(\Delta z_{k+1} + \Delta z_k\right) \qquad \Delta z_k^{(-)} = \frac{1}{2}\left(\Delta z_k + \Delta z_{k-1}\right).$$

Finally, the density averages ${\rho_r}^{(+/-)}$ at z-boundaries of a grid cell are given by

$$\rho_r^{(+)} = \frac{\left(\rho_{r,k+1} + \rho_{r,k}\right)/2 - \rho_0}{\rho_0} \qquad \rho_r^{(-)} = \frac{\left(\rho_{r,k} + \rho_{r,k-1}\right)/2 - \rho_0}{\rho_0}.$$

After this discretization, the solution of Equation (2.9) is reduced to the solution of a linear function of the hydraulic heads $h_{i,j,k}$, h_{i-1}, h_{i+1}, h_{j-1}, h_{j+1}, h_{k-1} and h_{k+1}. The coefficients of this linear equation can be computed from inputs of the simulation. These inputs include the cell sizes, the hydraulic conductivities at cell centers, the specific storage values at cell centers, and the source term. I will now rearrange Equation (2.10) to obtain a concise formulation of the linear equation that serves as a guideline for the implementation of the discretization in the source code of SHEMAT-Suite:

$$\begin{aligned}
\underbrace{\frac{S_s}{\Delta t}}_{R}\left(h_{i,j,k} - \hat{h}_{i,j,k}\right) - W = {} & \underbrace{\frac{k_x^{(+)}}{\Delta x_i \Delta x_i^{(+)}}}_{E}\left(h_{i+1} - h_i\right) - \underbrace{\frac{k_x^{(-)}}{\Delta x_i \Delta x_i^{(-)}}}_{C}\left(h_i - h_{i-1}\right) \\
& + \underbrace{\frac{k_y^{(+)}}{\Delta y_j \Delta y_j^{(+)}}}_{F}\left(h_{j+1} - h_j\right) - \underbrace{\frac{k_y^{(-)}}{\Delta y_j \Delta y_j^{(-)}}}_{B}\left(h_j - h_{j-1}\right) \\
& + \underbrace{\frac{k_z^{(+)}}{\Delta z_k \Delta z_k^{(+)}}}_{G}\left(h_{k+1} - h_k\right) - \underbrace{\frac{k_z^{(-)}}{\Delta z_k \Delta z_k^{(-)}}}_{A}\left(h_k - h_{k-1}\right) \\
& + \underbrace{\frac{k_z^{(+)}\rho_r^{(+)}}{\Delta z_k}}_{U} - \underbrace{\frac{k_z^{(-)}\rho_r^{(-)}}{\Delta z_k}}_{V}.
\end{aligned}$$

The capital letters denote the coefficients of the linear equation. With this notation the linear equation can be summarized as follows:

$$\begin{aligned}
-R\hat{h}_{i,j,k} - \left(U - V + W\right) = {} & A h_{k-1} + B h_{j-1} + C h_{i-1} \\
& + \left(\underbrace{A + B + C + E + F + G}_{D:=} - R\right) h \\
& + E h_{i+1} + F h_{j+1} + G h_{k+1}.
\end{aligned} \tag{2.11}$$

In SHEMAT-Suite, this linear equation is submitted to the linear solver.

To complete this section, a short discussion of the implementation of the finite difference dis-

cretization in SHEMAT-Suite is given. Figure 2.2 gives an overview of this implementation. The discretization is called from the time-stepping that implements the Picard iteration, and after the discretization the linear solver is called. Four subroutines set the parameters of the discretized linear equation. First, the coefficients A, B, C, D, E, F, G of the hydraulic heads in the linear equation are set. Secondly, sources and sinks are set. Thirdly, the constants of the linear equations are set. Finally, boundary conditions are imposed if cell i, j, k is a boundary cell. In SHEMAT-Suite, the parameters are set in so-called property modules. Inside a property module, a user can define a dependency of a parameter on dynamic variables, for example the dependency of the fluid density on pressure and temperature. In Figure 2.2 a complete list of the parameters set in the property module is given for each group of coefficients of Equation (2.11).

2.2 Kalman filtering

The ensemble Kalman filter (EnKF) is an ensemble-based data assimilation algorithm derived from the classical Kalman filter. The Kalman filter sequentially updates a state vector and its covariance matrix by optimally weighting model predictions and measurements. The Kalman filter provides optimal solutions for linear systems and Gaussian statistics.

2.2.1 Kalman filter

Before the EnKF is presented later in this section, the equations of the Kalman filter are recalled (Kalman, 1960, Kalman and Bucy, 1961). For the background on probability theory needed for this section, see Papoulis and Pillai (2002), and Sivia and Skilling (2006) for a Bayesian view on data assimilation. Background on filtering theory and early accounts of Kalman filtering can be found in Jazwinski (1970) and Gelb (1974).

Both the EnKF and the classical Kalman filter are originally methods for data assimilation of a dynamic state vector. The goal of these data assimilation methods is twofold: first, to compute the time variation of the dynamic variables in the state vector, and secondly, to update the dynamic variables according to measurement information. For filters, such as the EnKF or the classical Kalman filter the state vector is updated multiple times during the forward computation, whenever new measurement information becomes available. General data assimilation is computationally expensive and cannot be reduced to a simple set of equations. The value of the Kalman filter is that it applies a Gaussian assumption to the state vector of dynamic variables, as well as to the observations. In the classical Kalman filter, the forward equation is assumed to be linear, while the EnKF is also robust for non-linear forward equations. With its assumptions, the Kalman filter reduces the optimal data assimilation problem to a set of simple equations for the forward computation and update equation. These equations will be presented in the remainder of this section.

The classical Kalman filter is designed for data assimilation manipulating a multivariate Gaussian random vector. The multivariate Gaussian random vector is fully determined by a mean

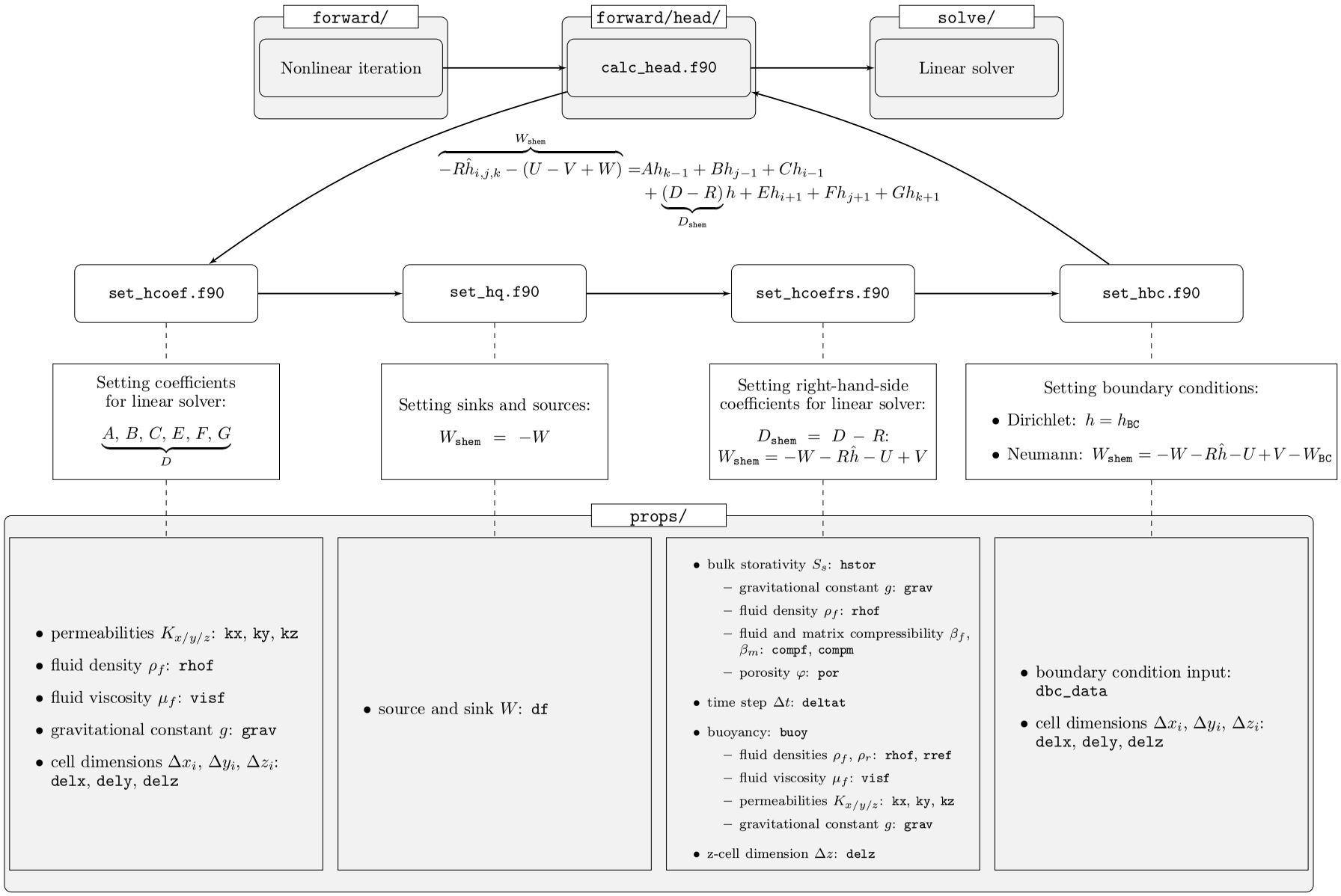

Figure 2.2: Implementation of the finite difference discretization in SHEMAT-Suite. The names of the functions in the SHEMAT-Suite property modules are given alongside the parameter names used in this thesis.

state vector and a covariance matrix. After the initial definition of mean state vector and covariance matrix is given, the two statistical moments are manipulated by the filter. In a forward computation, state vector and covariance matrix are propagated to the next observation time. The forward computation is carried out according to the partial differential equations of the dynamical system. Secondly, in an update equation, state vector and covariance matrix are modified according to available observation information. In the following paragraphs, forward and update equations of the Kalman filter are presented. The chapter is concluded by explaining how the Kalman filter is used for parameter estimation.

Let $n_s \in \mathbb{N}$ be the size of the mean state vector $\mathbf{x} \in \mathbb{R}^{n_s}$ of the Kalman filter. Then, the Kalman filter forward equation for the mean state vector is given by

$$\mathbf{x}^f = \mathbf{M}\mathbf{x}^0 \ \in \mathbb{R}^{n_s}, \tag{2.12}$$

where $\mathbf{x}^0 \in \mathbb{R}^{n_s}$ is the initial mean state vector. $\mathbf{x}^0$ is one of the inputs of the Kalman filter. For the classical Kalman filter, the state vector consists of mean values for dynamic random variables. The linear forward computation of the state vector is represented by the matrix $\mathbf{M} \in \mathbb{R}^{n_s \times n_s}$. The result of the forward computation is $\mathbf{x}^f \in \mathbb{R}^{n_s}$, called the mean prediction vector. Next to the forward equation for the mean state vector, the forward equation for the linear evolution of the covariance matrix is given by

$$\mathbf{P} = \mathbf{M}\mathbf{P}^0\mathbf{M}^T \ \in \mathbb{R}^{n_s \times n_s}, \tag{2.13}$$

where $\mathbf{P}^0 \in \mathbb{R}^{n_s \times n_s}$ is the initial covariance matrix. Here, the result is called $\mathbf{P} \in \mathbb{R}^{n_s \times n_s}$, the covariance matrix of the predicted state vector. Thus, the forward equations of the Kalman filter are the general equations of linear transformation for a Gaussian mean vector and covariance matrix. The predictions $\mathbf{x}^f$ and $\mathbf{P}$ serve as input to the update equation of the Kalman filter.

The Kalman filter update equation for the mean of the state vector is given by

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{P}\mathbf{H}^T \left(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}\right)^{-1} \left(\mathbf{d} - \mathbf{H}\mathbf{x}^f\right). \tag{2.14}$$

On the left-hand side of this equation, $\mathbf{x}^a \in \mathbb{R}^{n_s}$ is the state vector after the Kalman filter update. On the right-hand side of the equation, the computation of the Kalman filter update is detailed. The update equation is influenced by the statistical moments of the prediction computed by the forward computation: $\mathbf{x}^f$ and $\mathbf{P}$. The observation information that is processed by the filter is specified through $\mathbf{d} \in \mathbb{R}^{n_m}$, the mean vector of observations, where $n_m \in \mathbb{N}$ is the number of observations, and $\mathbf{R} \in \mathbb{R}^{n_m \times n_m}$ is the measurement error matrix. This means that the observations, like the state vector, are modeled as multivariate Gaussian random variables. Typically, the covariance matrix $\mathbf{R}$ can be specified as a diagonal matrix, if it can be assumed that there is no correlation between pairs of observations. Finally, $\mathbf{H} \in \mathbb{R}^{n_m \times n_s}$ is the measurement operator. The measurement operator maps the state vector to the space of observations. For this work, the linearity of $\mathbf{H}$ is ensured, because the state vector contains all observed variables (e.g. Gelb, 1974).

It is instructive to discuss a few limiting cases of the Kalman filter update. The first evident case is $\mathbf{d} = \mathbf{H}\mathbf{x}^f$, i.e. the means of the measurements $\mathbf{d}$ are equal to the means of the simulated observations. In this case, the so-called innovation $\left(\mathbf{d} - \mathbf{H}\mathbf{x}^f\right)$ is zero, and, consequently, there is no update. Now, two limiting cases for the covariance matrices will be discussed. First, let $\mathbf{P} >> \mathbf{R}$. The greater sign for the matrices stands for the condition that the diagonal elements of $\mathbf{P}$ are much larger than those of $\mathbf{R}$. In this case, the innovation gets a lot of weight and for the updated mean vector holds $\mathbf{x}^a \approx \mathbf{d}$ for all entries that correspond to the observations. The other limiting case $\mathbf{P} << \mathbf{R}$ results in a very small Kalman gain matrix multiplying the innovation, and thus almost no update is applied. This is to be expected, since in this case the measurements are very uncertain compared to the simulations.

The update equation for the covariance matrix $\mathbf{P}$

$$\mathbf{P}^a = \mathbf{P} - \mathbf{P}\mathbf{H}^T \left(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}\right)^{-1} \mathbf{H}\mathbf{P} \tag{2.15}$$

completes the Kalman filter update. $\mathbf{P}^a \in \mathbb{R}^{n_s \times n_s}$ is the updated covariance matrix. The matrix $\mathbf{K} \in \mathbb{R}^{n_s \times n_m}$ with

$$\mathbf{K} = \mathbf{P}\mathbf{H}^T \left(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}\right)^{-1} \tag{2.16}$$

is essential to both update equations and is called the Kalman gain matrix. The Kalman gain matrix determines the strength of the Kalman update based on the covariance information of the random vectors of the state and observation. After the update equation, the next forward simulation is started and $\mathbf{x}^a$ and $\mathbf{P}^a$ take over the roles of $\mathbf{x}^0$ and $\mathbf{P}^0$.

In this work, I use the Kalman filter for parameter estimation. The simplest way to do this is by augmenting the state vector by the targeted parameters (Gelb, 1974) as follows:

$$\mathbf{x}^f{}_{\text{augmented}} = \begin{pmatrix} \mathbf{x}^f \\ \mathbf{p} \end{pmatrix} . \tag{2.17}$$

For the forward computation, the parameters are constant, and for the update equation, the parameters are updated according to their covariances with state vector variables and observations. In the remainder of this thesis, EnKF state vectors are assumed to contain both dynamic variables and parameters. For the Kalman filter, choosing correct initial covariance matrices is especially challenging. The initially chosen covariances are very important since they determine the update of the parameters entirely. Thus, choosing wrong covariances may render results useless. For the ensemble Kalman filter (EnKF), the Monte Carlo version of the Kalman filter, choosing these initial covariance matrices is still very important. However, the hard and decisive step of choosing an initial covariance matrix is somewhat mitigated by the ensemble nature of the EnKF. In the EnKF, the state variables are specified by an ensemble of values. Next to the significant influence of the initial parameter values and their spread, the ensemble of parameters develops correlations with dynamic variables during the forward computation according to the sensitivity of the dynamic variables to the parameters. This way the update not only reflects the input of the initial parameter ensemble that should still be carefully selected, but also the influence of the forward computation. The EnKF was applied for permeability estimation us-

ing augmented state vectors by Naevdal et al. (2005) and Hendricks Franssen and Kinzelbach (2008).

2.2.2 EnKF

The EnKF is the Monte Carlo version of the Kalman filter (Evensen, 1994, Burgers et al., 1998). While the Kalman filter can only be applied to linear dynamics, the EnKF also performs well for non-linear dynamics. Instead of computing the full model covariance matrix as in the Kalman filter, the covariance matrix is approximated from an ensemble of model simulations in the EnKF. This way, the EnKF can be applied to larger systems with many unknowns. For the EnKF to give meaningful results, the ensemble should capture the main uncertainty sources relevant for the model prediction. For groundwater flow and species transport, the most important uncertainty source is typically the rock permeability governing the hydraulic conductivity. In this thesis, the EnKF is used for parameter estimation, by augmenting the state vector by rock permeability.

Like the Kalman filter, the EnKF consists of two main steps: The forward simulation (the result is indicated by superscript f), and the update equation (result indicated by superscript a). During forward simulation, each realization is simulated using the physical model M. Compared to the Kalman filter, I formulate the forward equation in a more general form by explicitly including the simulation time. For the forward simulation that computes the state vector at simulation time t_{j+1} from the state vector at simulation time t_j , one can write

$$\mathbf{x}_i^f(t_{j+1}) = M\left(\mathbf{x}_i^a(t_j)\right) \qquad i \in \{1, \cdots, n_e\}, j \in \{1, \cdots, n_t \in \mathbb{N}\}. \tag{2.18}$$

The vector $\mathbf{x}_i^a \in \mathbb{R}^{n_s}$ holds the ith realization of states and parameters at a given simulation time. In Equation (2.18) this simulation time is t_j *after* simulation. In the special case $j = 1$, $\mathbf{x}_i^a(t_1)$ holds the ith realization of the initial states and parameters. For the parameter estimation case, n_s is the sum of the number of states and the number of parameters, $n_e \in \mathbb{N}$ is the number of ensemble members, and $n_t \in \mathbb{N}$ is the number of assimilation times. In this work, the model M represents the forward solution of a subset of the following set of equations: the groundwater flow equation for hydraulic head, the heat transport equation for temperature, or the species transport equation for the concentration of a solute. The dynamic variables for the equations to be solved are part of $\mathbf{x}_i^a(t_j)$. The parameters that are part of $\mathbf{x}_i^a$ are left constant by M at all simulation times. The vector $\mathbf{x}_i^f \in \mathbb{R}^{n_s}$ holds the ith realization of states and parameters at a given simulation time. In Equation (2.18) this simulation time is t_{j+1} *before* assimilation. For simplicity, the function arguments (simulation times) will usually be dropped.

Note the differences between Equation (2.18) for the EnKF and Equation (2.12) for the Kalman filter. First, the forward equation for the EnKF is more general due to explicitly including the simulation times. Secondly, there are n_e equations for the EnKF, compared to the single equation for the mean state vector for the Kalman filter. Thirdly, while the forward equation for the Kalman filter is linear, for the EnKF, the realizations can undergo non-linear forward evolution. Fourth, the n_e realizations from the EnKF will be used to compute sample covariances,

thus a covariance equation analogous to Equation (2.13) in the Kalman filter is not needed for the EnKF. Finally, the distributed forward computation has advantages for parameter estimation already mentioned at the end of the last section. The evolution of the dynamic variables in the state vector will result in correlations with the parameters in the state vector according to the sensitivity of the variables to the parameters. This way, in the EnKF, the choice of a prior covariance matrix between dynamic variables and parameters is complemented by the correlations that are implicit in the forward equation. These correlations that stem from sensitivities are an important foundation of the permeability updates estimated in this thesis.

I now turn to the ensemble of observations given by

$$\mathbf{d}_i = \mathbf{H}\mathbf{x}^{\text{meas}} + \epsilon_i \qquad i \in \{1, \cdots, n_e\}\,. \tag{2.19}$$

The means of measured observations $\mathbf{y} = \mathbf{H}\mathbf{x}^{\text{meas}} \in \mathbb{R}^{n_m}$ are perturbed by Gaussian noise $\epsilon_i \in \mathbb{R}^{n_m}$ with covariance matrix $\mathbf{R} \in \mathbb{R}^{n_m \times n_m}$. This perturbation is important for estimating the correct uncertainties in the EnKF (Burgers et al., 1998). As for the Kalman filter, $n_m \in \mathbb{N}$ is the number of measurements at the time step under consideration. In the synthetic setups of this thesis the observations will be tracer concentrations or hydraulic heads from a reference model. In general, measured values are connected to the states and parameters of the model through the measurement operator $\mathbf{H} \in \mathbb{R}^{n_m \times n_s}$. In this thesis, the observations are part of the state vector, making $\mathbf{H}$ a matrix consisting almost entirely of zeros, with the exception of n_m ones on the diagonal. As result of the measurement equation, I obtain a vector of perturbed measurements $\mathbf{d}_i \in \mathbb{R}^{n_m}$ for each realization.

The following update equation is of central importance in the EnKF:

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}\left(\mathbf{d}_i - \mathbf{H}\mathbf{x}_i^f\right), \qquad i \in \{1, \cdots, n_e\}\,. \tag{2.20}$$

For each realization, the prediction from the forward simulation $\mathbf{x}_i^f$ is compared to the perturbed measurement vector $\mathbf{d}_i$. Then, $\mathbf{x}_i^f$ is updated according to the Kalman gain matrix $\mathbf{K} \in \mathbb{R}^{n_s \times n_m}$, which is given by

$$\mathbf{K} = \mathbf{P}_e\mathbf{H}^T\left(\mathbf{H}\mathbf{P}_e\mathbf{H}^T + \mathbf{R}\right)^{-1}. \tag{2.21}$$

Here $\mathbf{P}_e \in \mathbb{R}^{n_s \times n_s}$ denotes the ensemble covariance matrix of states and parameters. As for the Kalman filter, $\mathbf{K}$ favors updates if the ensemble covariances $\mathbf{P}_e$ are large compared to the measurement uncertainty $\mathbf{R}$.

As for the forward equation, it is interesting to record the main differences between the EnKF and the Kalman filter, this time for the update equation. Instead of one update equation for the mean state vector and one update equation for the covariance matrix, in the EnKF, there are n_e update equations for the ensemble. The covariance matrices and, consequently, the Kalman gain are computed from the ensemble. The introduction of the ensemble opens the method to non-linear forward equations and much larger model sizes. However, it also introduces disadvantages. First, the ensemble of realizations and the derived quantities may suffer from sampling error. And, secondly, the assumption of Gaussianity that was explicit in the statistical moments of the

Kalman filter is now only implicit for the EnKF through the optimality of the update equation. On one hand, this is an advantage since the EnKF can be applied to complex problems without further preparation. On the other hand, the conditions for the EnKF to have explanatory power have to be checked, for example making sure that either the probability distributions are nearly Gaussian or some extension of the method is used that may mitigate effects of non-Gaussian probability distributions (e.g. Zhou et al., 2011).

Details on how the EnKF is implemented in this work can be found in Evensen (2003). In the numerical software SHEMAT-Suite that is used in this thesis, the formulation of the EnKF from Evensen (2003) is extended for joint state-parameter updating according to Hendricks Franssen and Kinzelbach (2008). Building on this work, the EnKF and a number of methods derived from the EnKF that are detailed in Chapter 4.2.1 are implemented in SHEMAT-Suite.

Übe, übe deine Kräfte, was dich jetzt Mühe kostet
wird endlich maschinenmäßig werden.

Sudelbuch J, Fundstelle nach Promies 339
Georg Christoph Lichtenberg

3

Ensemble Kalman filtering in the numerical software SHEMAT-Suite

In this chapter, the software SHEMAT-Suite will be presented, a numerical code for simulating flow, heat and species transport in porous media. In particular, I will focus on the implementation of EnKF methods of SHEMAT-Suite that is part of this thesis. The suite of EnKF methods that are implemented in SHEMAT-Suite is introduced, followed by some guidelines for the scientific software development of SHEMAT-Suite during this thesis. Additionally, an introduction will be given to the motivation and significance of other parts of SHEMAT-Suite that existed prior to this thesis and were extensively refactored during the work on this thesis. A description of SHEMAT-Suite functionalities follows. The architecture that served as a guideline for the refactoring during this thesis is presented. The Theis pumping model serves as an illustrative example for how to use SHEMAT-Suite. Additionally, two simple inversion examples are described. After the impact of SHEMAT-Suite on the geosciences is detailed, the chapter is concluded by a summary of the guidelines for scientific software development that follow from the work on refactoring SHEMAT-Suite during this thesis and an outlook on how these guidelines may help the future development of SHEMAT-Suite and of other scientific software.

Contents

adapted from | Keller, J., Rath V., Bruckmann, J., Mottaghy, D., Clauser, C., Wolf, A., Seidler, R., Bücker, H.-M., Klitzsch, N. SHEMAT-Suite: an open-source code for simulating flow, heat and species transport in porous media, SoftwareX, 12, 100533 (2020).

3.1 Introduction

3.1.1 Motivation and significance of SHEMAT-Suite

Numerical simulations of coupled subsurface fluid flow and heat transport problems are routinely performed in a number of geoscientific fields, including geothermal energy, groundwater, nuclear waste storage and CO_2 sequestration. Simulations are increasingly based on large and complex geological models, requiring sophisticated and parallelized software packages, ideally with a history of extensive benchmarking and testing on real-life applications. SHEMAT-Suite is such a software: a versatile, general-purpose open-source code for simulating flow, heat and species transport in geological reservoirs. Conceptually, SHEMAT-Suite treats geological reservoirs as porous media. The code solves transient or steady-state, forward and inverse coupled problems in 1D, 2D, and 3D. It can handle a wide range of time scales and, thus, can address both technical and geological processes. SHEMAT-Suite can handle the non-linearities resulting from the variation of rock and fluid properties with temperature, pressure and species concentration. In addition, SHEMAT-Suite is capable of deterministic and stochastic inverse simulations and data assimilation, which enable parameter estimation and uncertainty quantification. These are important features, since subsurface flow and transport modeling is subject to data scarcity and large uncertainties of different sources.

For completeness, the main capabilities of the open-source SHEMAT-Suite are now shortly reiterated, a complete list of capabilities of open-source and closed-source SHEMAT-Suite can be found in the introduction. These functionalities of SHEMAT-Suite predate the work of this thesis, but everything that has to do with forward modeling and data assimilation was thoroughly refactored before publishing it as open-source code. As already seen in the introduction, the capabilities of the open-source SHEMAT-Suite include (1) a parallelization scheme for SHEMAT-Suite (Wolf,

2011); (2) the solution of the systems of coupled partial differential equations by implicit Picard iteration in a finite difference discretization (Lynch, 2005, Huyakorn, 2012); (3) a direct linear solver for small simulations, an implementation of the Bi-Conjugate Gradients Stabilized (Bi-CGStab) method for large simulations (van der Vorst, 1992); (4) additional solvers and preconditioners easy to integrate in SHEMAT-Suite owing to its modular structure of the solver implementation; (5) linear algebra functions used from the libraries BLAS (Lawson et al., 1979, Dongarra et al., 1988, 1990) and LAPACK (Anderson et al., 1999); (6) output of variable and parameter fields possible in many formats, most notably VTK (Schroeder et al., 2006) and HDF5[1] (The HDF Group, 1997-2022); (7) deterministic Bayesian inversion (Rath et al., 2006); (8) automatic differentiation of the forward code by the software Tapenade (Hascoët and Pascual, 2013) for computing the Jacobian in the deterministic inversion framework; (9) sequential Gaussian simulation (SGSim[2]) from the geostatistical library GSLib (Deutsch and Journel, 1995) for random field generation in the stochastic simulation mode within a Monte Carlo framework; (10) Bayesian updates implemented as ensemble Kalman filter (EnKF) analysis equations (Evensen, 2003). The part of SHEMAT-Suite that contains the EnKF implementation is discussed in this chapter in more detail.

The workflow of a typical application of SHEMAT-Suite proceeds in three steps: input, execution and output. Typically, the input step is the most time-consuming since it includes specifying the geometry of the model domain. The discretizations of time and space are specified in the input file along with the initial and boundary conditions for the dynamic variables (pressure or hydraulic reference head, temperature, dissolved species concentrations). Values of the required physical properties (e.g. porosity, permeability, thermal conductivity) are assigned to specific property units that represent model zones or rock-types with constant parameter values. Inverse simulations and parameter estimation require additional input files for specifying prior parameter values and measured data. For the execution step, the user should check the computing architecture in order to optimally use the parallelization of SHEMAT-Suite (Wolf, 2011). Finally, the type of output that is specified in the input file may be further processed and visualized. There are many output choices ranging from entire parameter fields at specified output times and time series at specified locations to inversion output such as the overall standard deviation of a parameter field. SHEMAT-Suite output files can be visualized and analyzed directly by using software such as ParaView (Ahrens et al., 2005) or Tecplot[3].

3.1.2 Suite of EnKF methods

The implementation of EnKF methods in SHEMAT-Suite was part of this thesis. The implementation uses as the starting point the EnKF update code by Geir Evensen[4]. Around this Fortran implementation of the EnKF update, a number of Fortran subroutines is implemented in SHEMAT-Suite alongside an easily manipulable input file and various choices for

[1] https://www.hdfgroup.org/downloads/hdf5/ [Accessed 2020/02/19]
[2] http://www.gslib.com/gslib_help/sgsim.html [Accessed 2020/02/26]
[3] https://www.tecplot.com/ [Accessed 2020/03/04]
[4] https://github.com/geirev/EnKF_analysis [Accessed 2020/02/19]

output. The EnKF methods implemented in SHEMAT-Suite include damping of the EnKF (Hendricks Franssen and Kinzelbach, 2008), the local EnKF (Hamill et al., 2001), the hybrid EnKF (Hamill and Snyder, 2000), the dual EnKF (Moradkhani et al., 2005, El Gharamti et al., 2013), the iterative EnKF (Sakov et al., 2012), the normal score EnKF (Zhou et al., 2011, Schöniger et al., 2012, Li et al., 2012), and the pilot point ensemble Kalman filter from Chapter 5. More details on these algorithms are given in Section 4.2.1. In this chapter, the structure of the implementation of these methods will be discussed. The implementation of the EnKF methods in SHEMAT-Suite follows two guidelines: first it should be relatively easy to add new EnKF methods, second a robust comparison between EnKF methods should be possible. For the robust comparison to be possible, two main measures are taken: as much of the source code as possible is reused among the various EnKF methods and source code optimizations, such as serial performance optimizations and parallel execution using OpenMP, make the large numbers of simulations possible that are needed for method comparisons.

3.1.3 Scientific software workflow

Some guidelines that provided the background of the scientific software development of SHEMAT-Suite during this thesis are now presented. Developing scientific software is different from developing commercial software in a number of respects, one being that the scientist is often the user and the developer of the software simultaneously (Segal and Morris, 2008). From this it follows that, in the beginning, the developer often does not know what the final software will look like. For commercial software, the implemented functionality is usually well-known from the beginning of the software development. Best practices for scientific software development should honor these conditions. One possible conclusion is that scientific software should be built and checked like a physical apparatus, for example a part of an experiment. This conclusion alongside a full list of best practices for scientific software development can be found in Wilson et al. (2014). In the geosciences, one main focus is often laid on the reusability of data, see for example the report by Radosavljevic et al. (2019). Additionally, there are scientific software packages that are developed and thereby following and setting best practices in the geosciences (Gerke et al., 2018, Hwang et al., 2017). There are also such scientific software packages for groundwater flow simulation. Examples are the alternatives to SHEMAT-Suite cited in the introduction, in particular the open-source variants.

While many scientifically developed software projects exist, I claim that a number of interesting and novel propositions can be deduced from the refactoring of SHEMAT-Suite. In particular, SHEMAT-Suite is interesting by its multi-functionality, including the core functionality of forward computation, but also stochastic inverse methods and deterministic Bayesian inversion using automatic differentiation. Multi-functionality is typical for scientific software that is developed as part of a number of scientific projects. I address this multi-functionality by proposing a specific workflow based on the version control software git. I employ git branches for managing the different functionalities of the overall software. This is, to the knowledge of the author, a

novel approach promoting git branches from the role of passive development helpers to an active role as a tool for scientific software development.

Now, some scientific guiding principles will be stated and it will be detailed how they have guided the software development of SHEMAT-Suite.

- **falsifiability**: There is an open-source **suite of test models**.
- **reproducibility**: Open-source availability of version-controlled **source code**.
- **Communication** of results: **Documentation**, both in the code (Doxygen[5]) and around it (a Wiki including a tutorial for new users)
- Logically sound **foundation for new developments**: Software development guidelines including **git branches** for functionalities and **modular**, easily adaptable source code.

The two basic principles falsifiability and reproducibility are met by ensuring that the source code and a suite of test models is available open-source. Reproducibility is closely connected to the last two principles: the communication and the foundation for new developments. By communication, I mean that a complete scientific work should not only be available, but accessible to fellow scientists. Therefore, the documentation of the code tries to address scientists, and the goal is not to distinguish between scientists and software developers. Finally, the code is written in a modular way. Like the git branches, this modular workflow supports the simultaneous development of functionalities, while these developments are still taking place in a single source code. In the subsequent part of this chapter, I will show how the presented principles are implemented for SHEMAT-Suite.

3.2 Software description

There are three compilation modes for SHEMAT-Suite: a forward mode for pure forward computation, an automatic differentiation mode (AD-mode) for deterministic inverse computation, and a stochastic mode for geostatistical simulation, stochastic data assimilation and parameter estimation. The forward mode has been refactored during this thesis and the stochastic mode was completely rewritten adding the EnKF methods for this thesis. Additionally, the structure of the code that is described in the following was imposed during this thesis.

Most of the user interaction with SHEMAT-Suite is restricted to the input file. For the forward mode, only one main input file is needed. For the inverse modes, additional input needs to be added in the main input file as well as in separate input files. Examples of SHEMAT-Suite input files are presented in Section 3.3, extracts from the input files can be found in Appendix A.

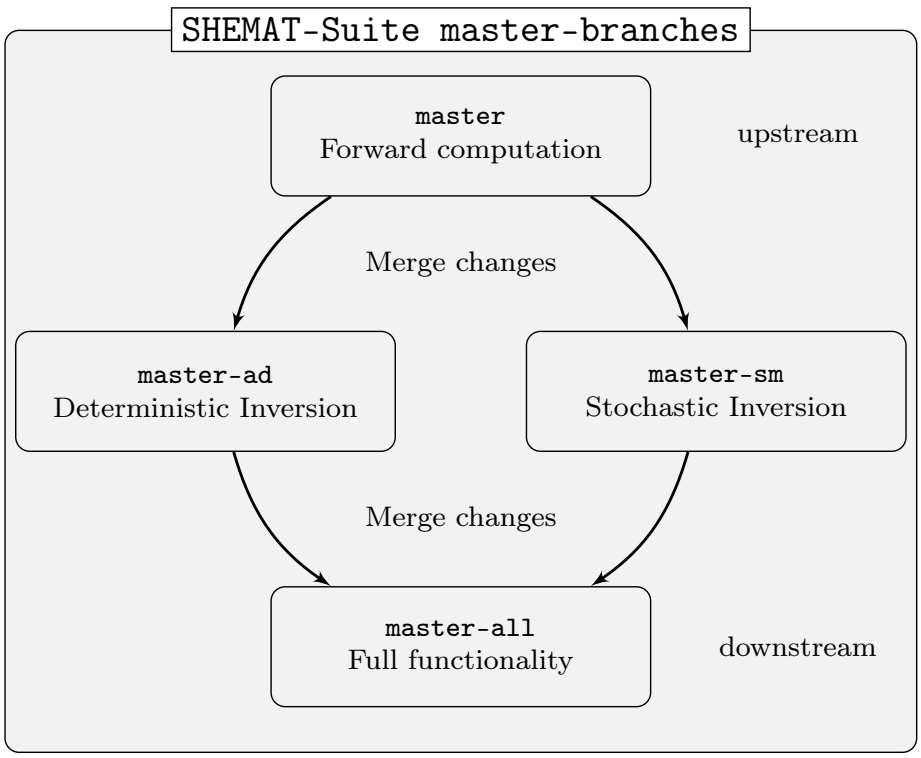

Figure 3.1: The branches of SHEMAT-Suite and their dependencies in the git-repository.

3.2.1 Software functionalities

The source code of SHEMAT-Suite is organized in two levels, the branch level and the directory level. On the branch level, the version control software git is used to divide the source code into branches. As usual in software development, git branches are used to develop the software, in particular the so-called 'master branch'. For SHEMAT-Suite, this concept is slightly extended to four master branches (Figure 3.1). The four master branches are derived from the three core functionalities of the software: (1) forward mode for the solution of flow, heat and dissolved species transport, (2) AD-mode (`ad`) for Bayesian inversion employing a derivative-based optimization scheme using tangent linear and adjoint codes generated via automatic differentiation (AD) for calculating exact Jacobian matrices, and (3) stochastic mode (`sm`) for Monte Carlo simulations, data assimilation and parameter estimation using the EnKF. The first of the four branches, `master`, can be compiled only in forward mode and is optimized for pure forward computations. The second branch, `master-ad`, can be compiled in forward mode and AD-mode. The third branch, `master-sm`, can be compiled in forward mode and stochastic mode. The fourth and final branch, `master-all`, can be compiled in all three modes. In general, a user can always work with `master-all`. However, the smaller codes may operate at better performance for the specific task, for example needing less memory or computation time.

Because of the special branch structure, it is important for SHEMAT-Suite developers to introduce changes in the correct branch and, subsequently, merge these changes into downstream

[5]http://doxygen.nl/ [Accessed 2020/05/11]

branches. Changes affecting the forward computations, should be added to `master`, changes affecting only the AD/inverse computations in `master-ad`, changes affecting only the Monte Carlo framework and stochastic computations in `master-sm`. The branch `master-all` should in principle only receive the changes to the other branches through merging. Rarely, specific changes may be needed to render the changes introduced from the upstream branches compatible.

Open-source documentation is available alongside the source code of SHEMAT-Suite. In the repository of SHEMAT-Suite, a Wiki[6] includes (1) general information about SHEMAT-Suite, (2) a tutorial for first time users, and (3) instructions on the use of the three main functionalities, forward computation, stochastic inversion, and deterministic inversion. For developers, a Doxygen documentation can be generated from source code headers. Additionally, the open-source git-repository `SHEMAT-Suite_Models-open`[7] contains test models for the various functionalities of SHEMAT-Suite and a Python script for testing simulation results against results of previous code versions. The `README` of `SHEMAT-Suite_Models-open` features an up-to-date list of test models and, if available, their analytic benchmarks. Scripts for pre- and post-processing are provided in the git-repository `SHEMAT-Suite_Scripts-open`[8]. For example, this repository contains a Python script for converting ASCII input files into HDF5 files that can be read by SHEMAT-Suite. Finally, the plan for SHEMAT-Suite is to gradually switch to continuous integration methodologies of software development. This will simplify the code development workflow, minimize the introduction of errors and thus assure the software quality.

3.2.2 Software architecture

This section explains the source code directories in the branches described in the previous section (see Figure 3.2 for a sketch of the source code directories used for forward computations). The `master` branch contains the following directories:

`/forward/` contains most of the code for forward computation, including the handling of input and output, the time discretization loop and the non-linear iteration loop. This code is parallelized using the shared-memory parallel programming paradigm OpenMP aiming for multi-core workstations.

`/solve/` contains the solution of systems of linear equations. The separation from `/forward/` ensures the easy implementation of additional solvers and preconditioners and is crucial for the usage of an automatic differentiation software.

`/props/` contains property modules defining the dynamic behavior and coupling of fluid and rock properties. Property modules can be designed for using both constant, or pressure- and temperature-dependent fluid and rock properties. Additionally, the influence of dissolved species concentrations, typically salt concentration, on fluid properties such as density can be accounted for. The modular structure allows for adding site- or application-specific property relationships.

[6] https://git.rwth-aachen.de/SHEMAT-Suite/SHEMAT-Suite-open/-/wikis/tutorial [Accessed 2020/03/04]
[7] https://git.rwth-aachen.de/SHEMAT-Suite/SHEMAT-Suite_Models-open/ [Accessed 2020/03/16]
[8] https://git.rwth-aachen.de/SHEMAT-Suite/shemat-suite_scripts-open [Accessed 2020/03/17]

`/user/` contains files for user-defined input subroutines, output subroutines, a user-defined Fortran-module, and a general user-defined subroutine called at the beginning of Picard iterations. The complete set of these subroutines is called a user module of SHEMAT-Suite. User modules provide an easy and flexible way for adding and adapting problem-specific functionalities. The open-source SHEMAT-Suite provides an exemplary template user module in the directory `none`. The directory `none` contains subroutine templates without functionality that serve as a starting point for developing new user modules.

`/hdf5/` contains interfaces with the HDF5 library for input and output.

`/cmake/` contains compilation utilities using CMake tools[9].

`/doc/` contains input for generating the Doxygen documentation.

The branches `master-sm` and `master-ad` add the following directories.

`/simul/` (`sm`) contains subroutines for geostatistical simulation, Monte Carlo computation and EnKF updates.

`/inverse/` (`ad`) contains the deterministic inversion code.

`/mkAD/` (`ad`) contains scripts for generating AD-code.

3.2.3 Implementation of EnKF methods

The EnKF methods of SHEMAT-Suite are implemented in the directory `/simul/` of the branch `master-sm`. There are two core components of the implementation of EnKF methods: first the code for the EnKF update[10], and secondly, the wrapper routine `enkf_iter` that calls Fortran subroutines for all the major parts of the EnKF implementation. The file `enkf_iter.f90` can be found in the open-source SHEMAT-Suite[11]. The implementation of the EnKF methods can be best explained by analyzing the structure of `enkf_iter`. This will be done in two steps: first, the general order of functions in the file, secondly, the locations of the manipulations for the EnKF methods will be presented.

At the center of the wrapper routine `enkf_iter`, there is the main EnKF loop. I will first discuss the steps of the code outside of the main EnKF loop. Before the loop is started, there are two major steps in `enkf_iter`: first, there are the input routines, second there are the SGSim routines (Deutsch and Journel, 1995). The input routines consist of a subroutine that searches the ASCII input file of the EnKF for keywords and reads the input. Additionally, there is a subroutine for checking the inputs. In particular, it is checked that numerical inputs are in physically and logically allowed ranges. Finally, arrays that are specific to the EnKF are allocated according to the input. Following the input routines, the sequential Gaussian simulation (SGSim) is carried out, initializing parameter fields according to input files defined in the geostatistical library GSLib

[9] https://cmake.org/ [Accessed 2020/02/26]
[10] https://github.com/geirev/EnKF_analysis [Accessed 2020/02/19]
[11] https://git.rwth-aachen.de/SHEMAT-Suite/SHEMAT-Suite-open/ [Accessed 2020/05/10]

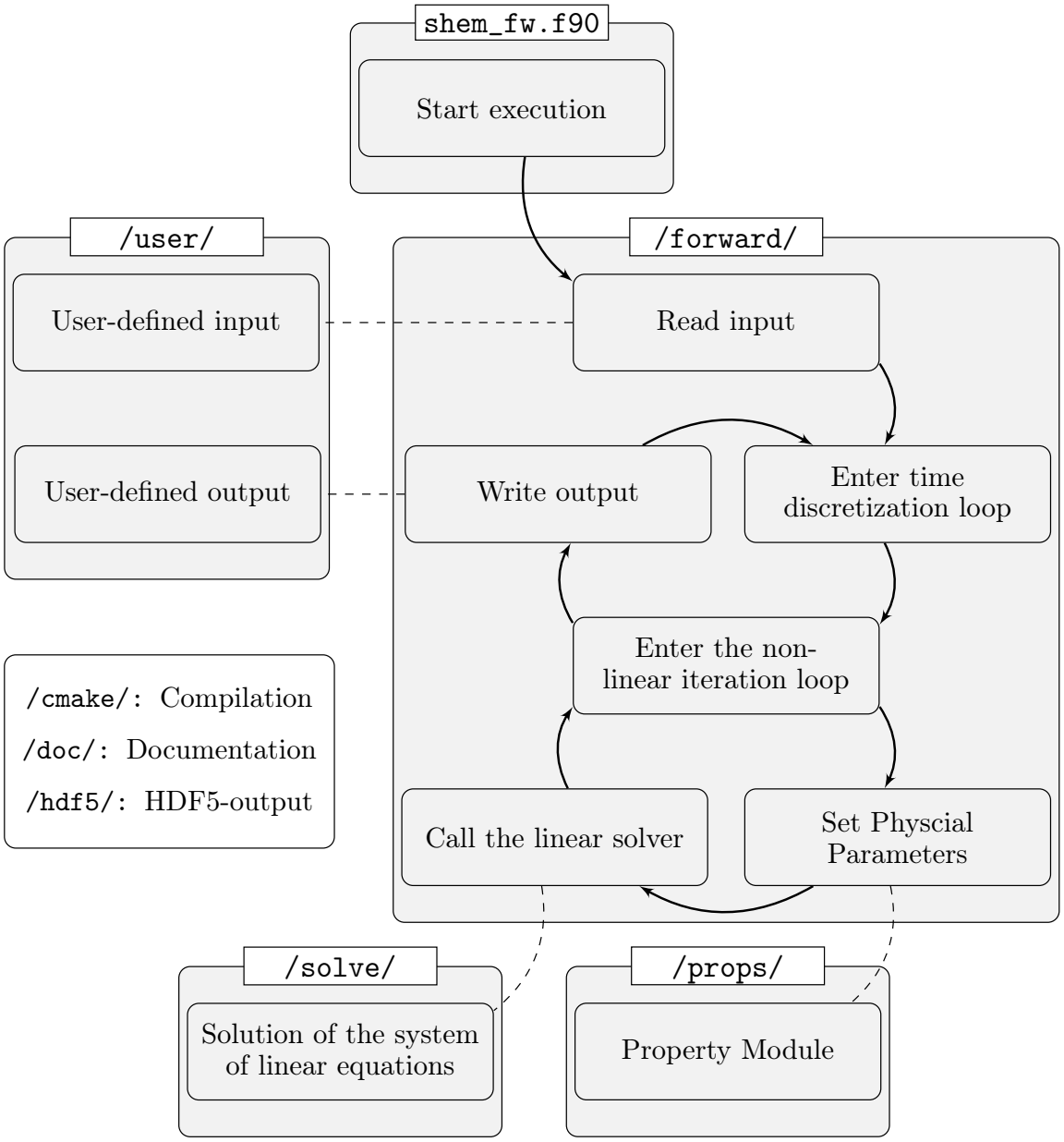

Figure 3.2: Diagram of the forward computation in SHEMAT-Suite highlighting the role of various source code directories.

(Deutsch and Journel, 1995). This also includes setting the random seeds for all the important random number generations. The main EnKF loop that will be discussed in the next paragraph follows. At the end of the EnKF loop, an updated parameter field has been generated. After the main EnKF loop, there are two steps remaining: first, final output of the EnKF run can be written. Next to the full variable and parameter fields, this includes derived output such as the root-mean-square errors or overall standard deviations that are used in the work of the following chapters of this thesis. At the end of `enkf_iter`, the option for so-called postcomputation is available. This postcomputation consists of a pure forward computation using the updated parameter fields.

The main part of `enkf_iter`, the EnKF loop, is now described in more detail. The loop iterates over the number of parameter updates of the EnKF, i.e. the times at which observation information is available. The principle of implementation of the EnKF is explained by outlining the essential steps of a single iteration. First, the forward computation from the previous to the current observation time is carried out by calling the subroutines in the general forward code of SHEMAT-Suite. After this, the EnKF state vector is generated and filled with the forecasted values. In each iteration, writing output files is possible. Output is written both, before the EnKF update, and after the update. In between, the EnKF update is carried out by calling the Fortran routines from Evensen (2003). A lot of the output written before and after update is similar. This way, the updates of dynamic variable fields and parameter fields can be monitored closely if needed.

It remains to describe, how the suite of EnKF methods is implemented. By implementing these methods in a modular way, specific routines can be called for manipulating the state vector or covariance matrices, while at the same time a large amount of the code is reused for all methods. The EnKF methods can be grouped into three classes according to their specific implementation: first, manipulation of the state vector at the beginning of a loop iteration, secondly, manipulation of the EnKF update, and thirdly, manipulation of the state vector right before and after the EnKF update. The methods of the first group are the iterative EnKF and the dual EnKF. These two EnKF methods, as well as the others in this paragraph are explained in more detail in Section 4.2.1. For the iterative EnKF, mainly the counting of iterations has to be adapted, or restarted. For the dual EnKF, the counting is manipulated in order to double the number of iterations. Additionally, the state vector is manipulated to contain only parameters in the first iteration of a certain simulation time and only dynamic variables in the second iteration of the same simulation time. The second group of EnKF methods includes the damped EnKF, the local EnKF and the hybrid EnKF. These methods explicitly manipulate the EnKF update step. The damped EnKF is just a simple manipulation of the update, where a damping factor is multiplied to parameters of the EnKF update. In contrast, the local EnKF and the hybrid EnKF include a non-trivial manipulation of the EnKF update, in particular including manipulations of the covariance matrices in this update. Finally, the third and last group of manipulations includes the normal score EnKF and the pilot point EnKF. Here, the state vector of the EnKF is manipulated right before and right after the update. For the normal score EnKF, this manipulation consists of the normal score transform and its inverse. For the pilot

point EnKF, the state vector is restricted to the pilot points before the EnKF update and an interpolation to the full state vector is implemented after the update.

3.3 Illustrative examples

This section presents three simple example simulations for illustrating the SHEMAT-Suite's versatile simulation capabilities, corresponding to the three modes `fw`, `ad`, and `sm`. The examples for the modes `fw` and `ad` mode (Sections 3.3.1 and 3.3.2) are meant to show the general usage of SHEMAT-Suite and have, otherwise, no connection to the work in this thesis. The third example (Section 3.3.3) on the other hand introduces the input file for the permeability estimation that is used for comparing EnKF methods in the following chapters of this thesis. For additional examples of scientific applications of SHEMAT-Suite, the interested reader is referred to the literature cited in Sections 1.1 and 3.4 and in Table 3.2. The complete input files for the examples in this section and for additional examples illustrating the various functionalities of SHEMAT-Suite can be found in the test model repository `SHEMAT-Suite_Models-open`[12].

3.3.1 Forward computation

The Theis pumping model serves as illustration of SHEMAT-Suite, by comparing the analytical solution of the Theis pumping well model (Theis, 1935) to the numerical solution computed in SHEMAT-Suite. The Theis model is a solution of the flow equation, computing the transient water table drawdown resulting from a pumping well in a horizontal, confined aquifer. I compare the numerical solution by SHEMAT-Suite to the analytical solution

$$h(r,t) = h_0 - \frac{Q}{4\pi\,\Delta z\,k} \cdot w(\frac{r^2 S}{4\,\Delta z\,k\,t}). \tag{3.1}$$

Here, h [m] denotes the water table including drawdown at distance from the well r [m] and time t [s]. On the right-hand side of the equation, h_0 [m] denotes the constant water table without the influence of the pumping well. Q [$\mathrm{m^3\,s^{-1}}$] is the pumping rate of the well, k [$\mathrm{m\,s^{-1}}$] is the hydraulic conductivity of the fluid-matrix mixture, S [$-$] is the storage coefficient, and Δz [m] denotes the thickness of the aquifer. For this homogeneous aquifer, the specific storage coefficient S_s from the groundwater flow equation is given by

$$S_s = \frac{S}{\Delta z}. \tag{3.2}$$

The function w is called the well function, a table of function values can be found in Segol (1994).

The SHEMAT-Suite Input File A.1 for the Theis problem can be found in Appendix A. The general SHEMAT-Suite input file consists of a number of sections defining the subsurface model.

[12] https://git.rwth-aachen.de/SHEMAT-Suite/SHEMAT-Suite_Models-open/ [Accessed 2020/03/16]

The first section provides general information about the run, for example a title and the specification of the property module, in this case `const`. This property modules defines constant fluid parameters such as fluid density and viscosity as well as constant rock parameters such as permeability throughout the simulation. The second section defines the grid starting with the number of cells in x-y-z-direction and followed by the cell sizes. In this case, the grid is coarser away from the well and finer in the center of the model near the pumping well. Under time step control, a transient simulation is specified and some solver parameters influencing the time stepping are set to their default values. More information on these special parameters can be found in the SHEMAT-Suite documentation. The time unit is set to one day (86 400 s) and the simulation runs for 3.5 days evenly split into 100 time steps. For larger simulations, SHEMAT-Suite offers the possibility of dynamic time-stepping sensitive to the convergence speed. The precision and maximum number of iterations are set for the linear solver method Bi-CGStab. The lateral head Dirichlet boundary conditions are set to 20 m, and top and bottom of the model are no flow boundaries. In the center of the model, the pumping well is represented by a sink at cell (35, 35, 1). This sink is specified as a pumping rate of $-0.001\,\mathrm{m^3\,s^{-1}}$ resulting in the water table drawdown around this cell that can be compared to the analytical Theis formula. The initial hydraulic head is set to 20 m for all 4 761 grid cells. The properties of the fluid (water) and the rock matrix are set. For water, these properties include density ($1\,000\,\mathrm{kg\,m^{-3}}$), compressibility ($5 \cdot 10^{-8}\,\mathrm{m\,s^2\,kg^{-1}}$), and viscosity ($10^{-3}\,\mathrm{kg\,m^{-1}\,s^{-1}}$). For the rock matrix, the properties include porosity (10 %), permeability ($10^{-12}\,\mathrm{m^2}$), and compressibility ($10^{-8}\,\mathrm{m\,s^2\,kg^{-1}}$). Other parameter inputs include thermal parameters and parameters for species transport. Since this illustrative example is for groundwater flow only, these additional parameters do not influence the simulation. In the SHEMAT-Suite documentation, users find a complete walk-through of a number of example input files including this one for the Theis model. Finally, the output file types are set to VTK and HDF5.

The numerical results for the Theis model are shown in Figure 3.3. The hydraulic head field visualizes the drawdown around the pumping well after 3.5 days. The HDF5 output of SHEMAT-Suite is visualized using Python and matplotlib (Hunter, 2007). The comparison with the analytical Theis solution reveals that the absolute error is smaller than 10 cm, even in the center of the model, where the influence of the pumping well is greatest and in the order of meters. Because of this, the relative errors in the center of the model are smallest and less than 1 %. Near the boundaries, the relative errors are in the order of 1 % due to a negligible influence of the pumping well. The Theis model and other test models for SHEMAT-Suite are published in the `SHEMAT-Suite_Models-open` repository[13].

3.3.2 Inverse simulation

A simple two-dimensional model (a vertical cross-section of the subsurface) is used for illustrating an inverse simulation for deterministic parameter estimation with SHEMAT-Suite. The model domain is 150 m deep and 300 m wide with an equidistant discretization of 2 m. The model

[13] https://git.rwth-aachen.de/SHEMAT-Suite/SHEMAT-Suite_Models-open [Accessed 2020/03/16]

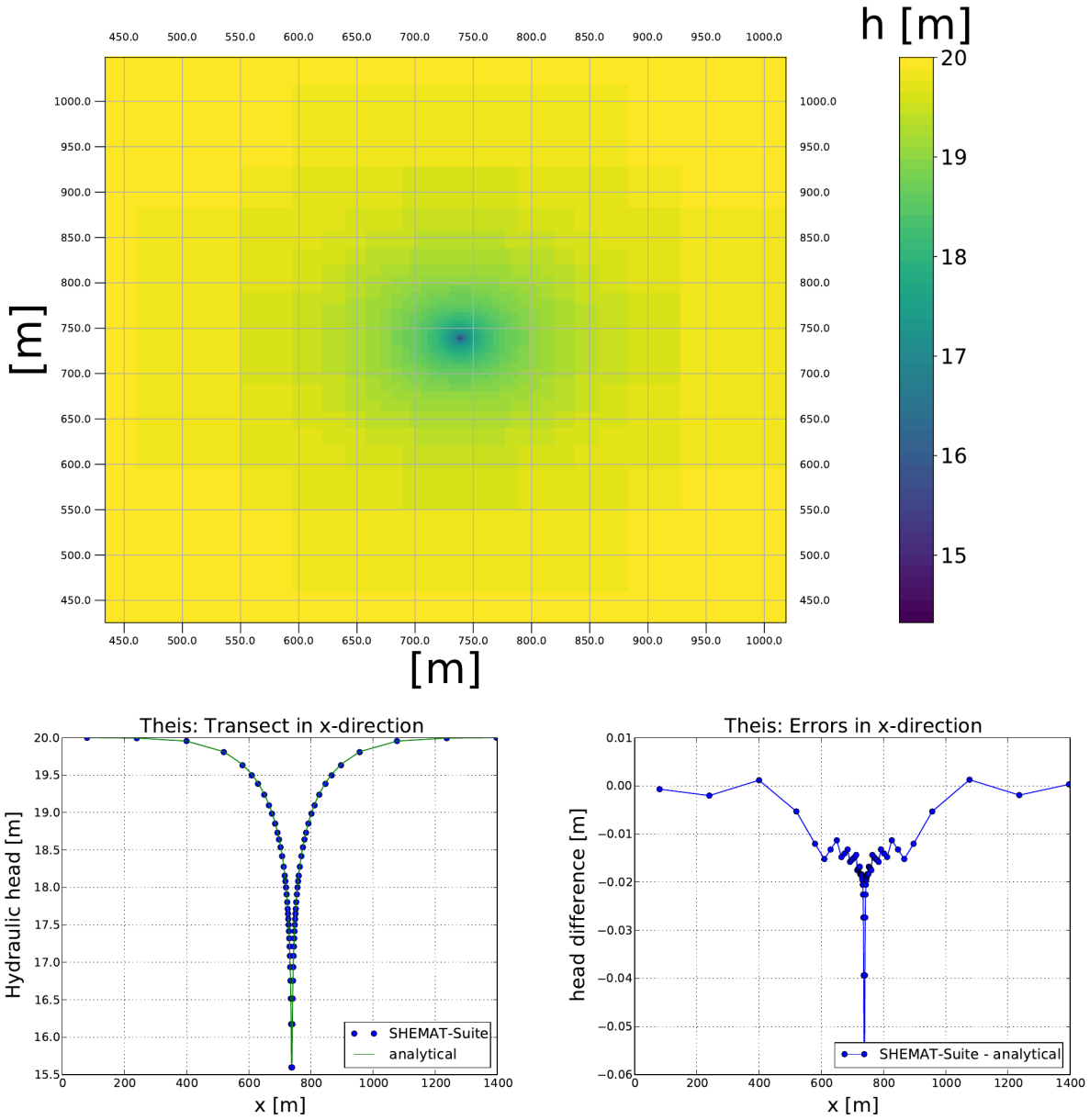

Figure 3.3: Results of the THEIS model. a) head distribution after 3.5 days, b) transect across the pumping well, and c) difference between the numerically and analytically calculated head along the transect.

consists of nine homogeneous model units. The first unit (`uindex 1`) represents a dummy unit for the air above the earth surface and units `uindex 2` to `9` represent rock layers which differ in porosity, permeability and thermal conductivity (Figure 3.4 a)).

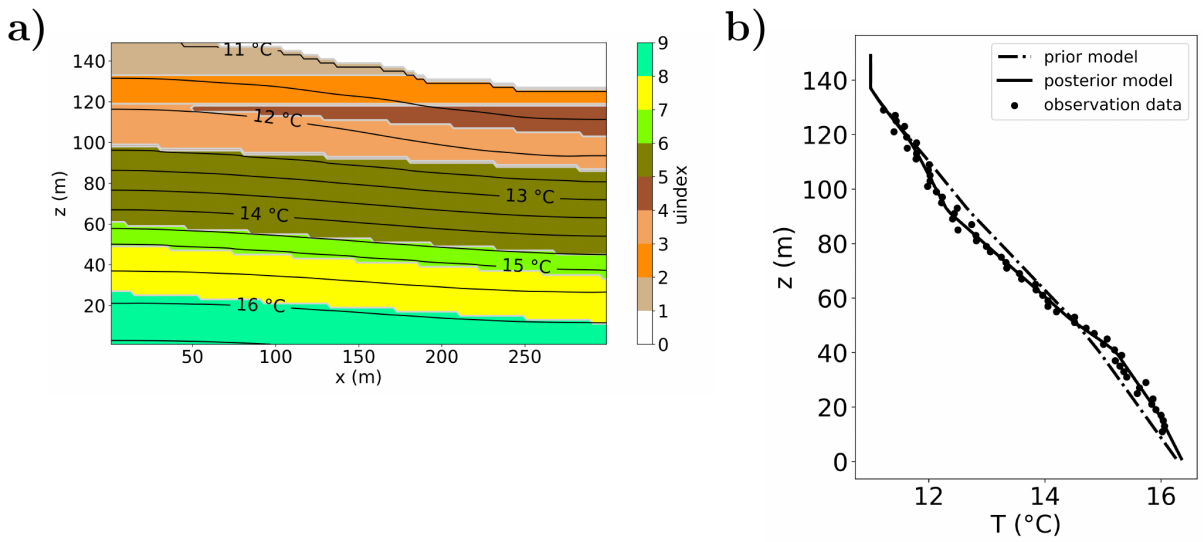

Figure 3.4: a) Cross-sectional model for illustrating an inverse simulation with SHEMAT-Suite. The white layer (`uindex 1`) represents the air above the topographic surface. Colored layers represent rock layers with different petrophysical properties specified in the input file. The black contour lines denote the temperature profile of the synthetic reference model. b) Observed temperature data (black circles), prior (dashed-dotted black line) and posterior (solid black line) temperature logs in the model center. The posterior model is the inversion result.

The inverse problem consists of determining the thermal conductivity of each of the six lower rock layers (i.e., `uindex 4` to `uindex 9`) using a temperature log available in the model center. In this synthetic example, it is assumed that all other rock properties in each layer and all boundary conditions are known without uncertainty. The prior model is specified in the input file similar to the forward problem input described in Section 3.3.1. Dirichlet boundary conditions are used for temperature and head along the topographic surface: the surface temperature is 11 °C and hydraulic head follows the topography. Therefore, these boundary conditions are defined in cell-by-cell mode in the input file (see bottom of Input File A.2 in Appendix A). A Neumann boundary condition defines the specific basal heat flow at the bottom of the model domain. The constant specific basal heat flow of 0.06 $\mathrm{W\,m^{-2}}$ can be defined using a simplified input specification (see Input File A.2 in the appendix, second to last keyword entry). This example simulation is a coupled simulation of flow and heat transport where fluid and rock properties are pressure- and temperature dependent. The parameter dependencies are contained in the property module `bas` that is specified in the input file following the keyword `PROPS`. For compatibility, SHEMAT-Suite has to be compiled using the compiler flag for the property module `bas` and using AD-mode for enabling its inversion capability. This capability is a prerequisite for `runmode 2` that is defined in the Input File A.2. For the pure forward code from the branch `master`, the `runmode` would have to be `0` or `1`. The prior rock properties are specified for each model unit below the input file

keyword `units`, analogous to a forward simulation. In this synthetic example, all rock properties are the same as in the true reference model, except for thermal conductivity of units `uindex 4` to `uindex 9`. There, thermal conductivity is assumed to be $2.0\,\mathrm{W\,m^{-1}\,K^{-1}}$ prior to inversion.

For the inverse mode, SHEMAT-Suite needs additional specifications. These additional keywords can be included in the main SHEMAT-Suite input file, or a second input file can be generated. Here, the additional keywords for the inverse simulation are included in the main input file. The part of the SHEMAT-Suite input file that is specific for inverse simulation is shown in Input File A.3 in the appendix. The keyword `inverse` is followed by some specifications for the inverse solver. The entries below the keywords `enable unit`, `enable property`, and `optimize property` define the model units and parameters that are enabled for inversion. Here, thermal conductivity is enabled for inversion in units `4` to 9. The entries below the keyword `errors` give the prior variances of the parameters to be estimated for each model unit. I assume a prior variance of $2.0\,\mathrm{W\,m^{-1}\,K^{1}}$ for each unit that is enabled for the inversion. Finally, I need to specify the observation data used for inversion. Temperature data are listed below the keyword `data`. The measurement error (0.1 K) and the grid indices of each data point are specified here as well.

The inverse simulation results in deterministic predictions of the thermal conductivity in units `uindex 4` to `uindex 9` and of the respective posterior model states (i.e., head and temperature). For this synthetic model, for which the true reference solution is known, the inversion result can be evaluated by comparing the true and posterior model directly. Table 3.1 lists the true reference and the posterior thermal conductivity for units `4` to `9` respectively. Figure 3.4 b) shows the temperature log in the model center for the prior model (forward run of the example model), the temperature data used, and the posterior temperature log after inversion. The prior model is not able to capture the observed temperature gradient, which is caused by advective and conductive heat transport. Inversion of the temperature data results in a good estimation of the thermal conductivity in units `4` to `9`. Maximum posterior deviation from temperature data is below the measurement error and the maximum posterior deviation of the estimated thermal conductivity is $0.42\ \mathrm{W\,m^{-1}\,K^{1}}$ (or 16.8 %) in unit `9`. Unit `9` is the only unit with a relative error of more than 10 %. This may be attributed to boundary effects at the lower boundary of the model. All in all, the inversion leads to a more realistic simulation of the observed temperature gradient.

Table 3.1: True reference thermal conductivity λ_{true} compared to posterior thermal conductivity $\lambda_{posterior}$ resulting from the inverse simulation for model units `4-9` (`uindex`). The assumed prior thermal conductivity is $2.0\,\mathrm{W\,m^{-1}\,K^{-1}}$ for each model unit.

`uindex`	λ_{true} ($\mathrm{W\,m^{-1}\,K^{-1}}$)	$\lambda_{posterior}$ ($\mathrm{W\,m^{-1}\,K^{-1}}$)	`Error (%)`
4	3.5	3.30	5.7
5	3.00	3.12	4.0
6	1.50	1.50	0.0
7	1.00	0.95	5.0
8	2.00	2.04	2.0
9	2.5	2.92	16.8

3.3.3 Stochastic simulation

For stochastic simulation using SHEMAT-Suite, in particular for data assimilation and parameter estimation using the ensemble Kalman filter (EnKF, Evensen, 2003), additional input is required compared to a forward computation. In this paragraph, the general input for a stochastic parameter estimation is discussed. Here, I discuss a synthetic parameter estimation of the type analyzed in the other chapters of this thesis. In such a synthetic stochastic parameter estimation, one forward simulation using a so-called synthetic true parameter field is computed and synthetic observations are recorded from this forward simulation. As a substitute for real observations, these synthetic observations are then used as input for a subsequent parameter estimation simulation, in which an ensemble of prior parameter realizations is subject to a given number of EnKF updates. The EnKF assimilation update the prior parameter realizations based on the differences between the synthetic observations and an ensemble of simulated observations. The simulated observations are calculated from forward simulations that in turn rely on the prior parameter realizations.

Here, an example of a synthetic stochastic parameter estimation called `TRACER` will be discussed. In this parameter estimation example, the EnKF will be used for estimating two-dimensional permeability fields from tracer concentration observations. Uniform salt concentrations at the south of the model start spreading to the north and are recorded at two observation locations in a forward computation using the synthetic true permeability field shown in Figure 3.5 a). The mean of the logarithms of the synthetic true permeabilities (in $[m^2]$) is -12.0. The ensemble of prior permeability fields will be generated using a different mean permeability of -12.5.

A number of input files are needed for the EnKF run:

- Forward mode input and additional stochastic mode input in the general SHEMAT-Suite input file `TRACER`.
- EnKF input in the SHEMAT-Suite EnKF input file `TRACER.enkf`
- Input for the sequential Gaussian simulation (SGSim) of an ensemble of prior permeability realizations (Deutsch and Journel, 1995)
- Values of the true permeability field for comparison with the estimated parameter values.
- Salt concentration data in a formatted observation file.

Please find more information in the Wiki of SHEMAT-Suite[14] on (1) the specialized input of the sequential Gaussian simulation, (2) the true permeability values which are stored as Tecplot files[15] and (3) the observation file. In this section, the SHEMAT-Suite input in `TRACER` and `TRACER.enkf` will be presented in more detail.

First, the additional input in the general SHEMAT-Suite input file compared to the forward mode is discussed. Input File A.4 in Appendix A shows the most important part of the input

[14] https://git.rwth-aachen.de/SHEMAT-Suite/SHEMAT-Suite-open/-/wikis/tutorial [Accessed 2020/03/04]
[15] https://www.tecplot.com/ [Accessed 2020/03/04]

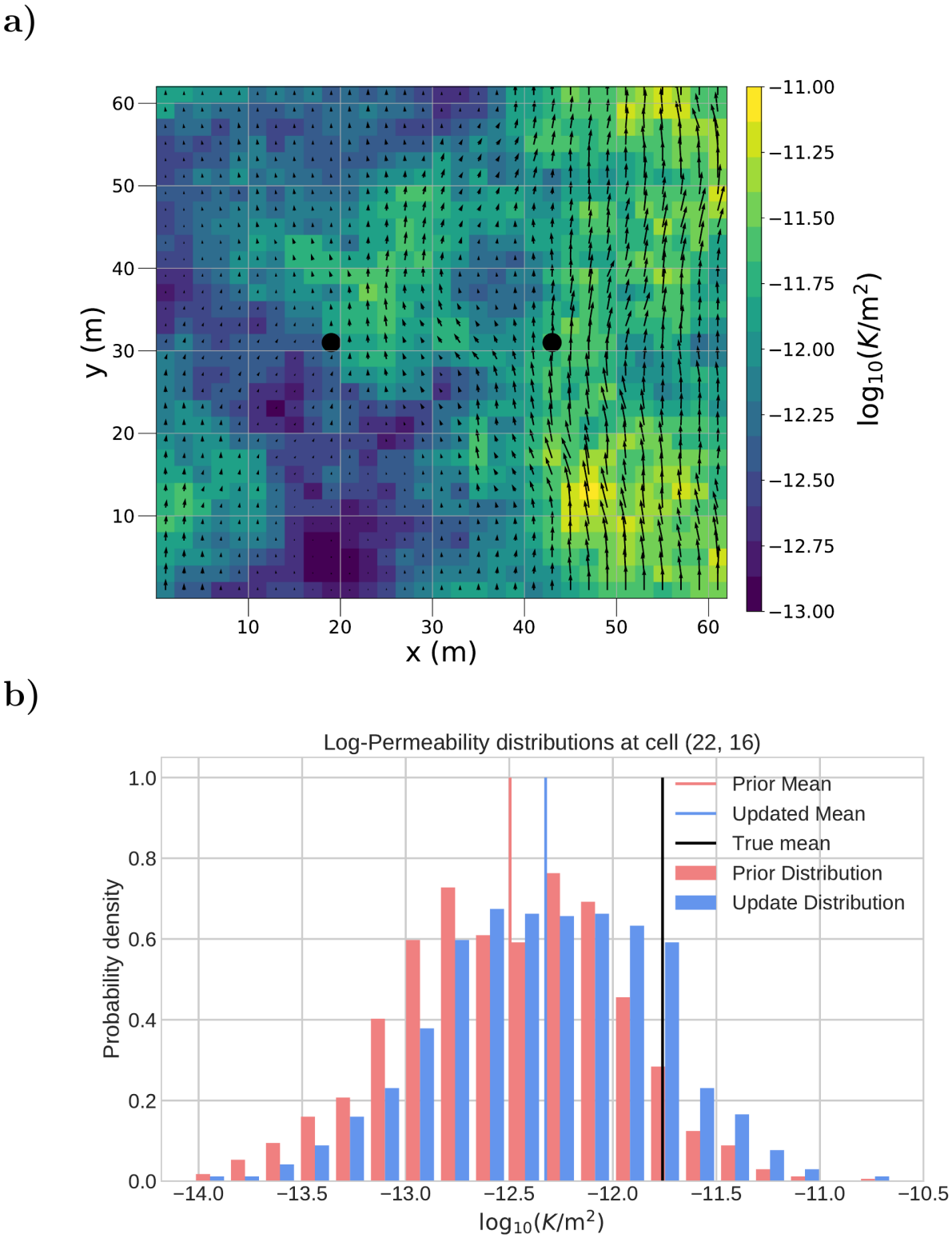

Figure 3.5: a) Synthetic true, horizontal permeability field (2-D model) with two observations for tracer concentrations (black dots). The same synthetic true permeability field is used for the synthetic parameter estimation `TRACER` from Chapter 4. b) Permeability distribution before and after the ten updates during the synthetic stochastic parameter estimation `TRACER`.

file for an EnKF simulation. Next to the title and the information output switches, the `runmode` has to be set to `3` for the stochastic mode using EnKF instead of `0` for the forward mode. Property module and user directory are set as in forward mode. The first input that is specific to the stochastic mode is `# simulate`. Here, the number of prior permeability field realizations is set to 1 000. Only one EnKF simulation is wanted in this parameter estimation. In special cases, one might want to restart the EnKF with the already updated parameter values. No postcomputation (forward computation using the updated permeabilities) is wanted, since I just want to analyze the updated permeability values. This is specified under `# enkf postcompute none`. Next to these general stochastic mode inputs, some important inputs for the Sequential Gaussian simulation have to be specified in the general SHEMAT-Suite input file `TRACER`. Under `# parameter group`, the SGSim input file is named, the permeability is chosen as the parameter to be simulated (it is the fourth parameter) and `log` means that the decimal logarithm of the permeability is used. This transformation is needed to comply better with Gaussian assumptions inherent to the EnKF. The input `# standard deviation` is a dummy input. The numbers are used internally for converting logarithmic permeabilities from the decimal logarithm to the natural logarithm. The rock parameters given under `# units` are all used as in forward mode, except for the permeability value. The permeability value is replaced by the simulated fields from the sequential Gaussian simulation. Finally, `# split units` ensures that each cell of a certain parameter unit becomes a unit in itself, i.e. each cell gets its own permeability value. This is a prerequisite for the sequential Gaussian simulation.

Now, the inputs in the SHEMAT-Suite EnKF input file are discussed as shown in Input File A.5 in Appendix A. The number of observation intervals, i.e. the number of EnKF updates is specified under `# nrobs_int`. Here, the input specifies 10 updates, so I run one EnKF simulation consisting of ten update times. Then, state vector contents are specified, choosing from the dynamic variables hydraulic head, temperature, species concentration, and from the parameters permeability, thermal conductivity and porosity. In this example, hydraulic head, tracer concentration and permeability are part of the state vector. The observation variances specify the measurement error, in this example for concentration observations: $50 \cdot 10^{-6}$ mol l^{-1}. The comparison with the true permeability field is wanted (`# checktrue`). Finally, the input file names for the truth and the observations are specified.

Several outputs are written by the EnKF run. Here, two outputs are discussed: permeability distributions at specific locations in the model domain and overall RMSE between the updated permeability fields and the synthetic true permeability field. Prior and updated permeability distributions at a location close to the eastern observation location are shown in Figure 3.5 b). Permeabilities are updated towards larger values, i.e. a more permeable subsurface. From the overall RMSE output, one can observe that the residual is reduced from 0.623002 to 0.440239. The residual values are differences of permeability exponents. The remaining difference between the true permeability field and the synthetic true permeability field can be explained by four reasons: (1) a smoothing of the permeability realizations in the updated mean permeability field, (2) the measurement uncertainty, (3) the sampling uncertainty from the ensemble size, and (4) the limited number of observations.

3.4 Impact

I use two aspects for describing the impact of SHEMAT-Suite: geosciences and computational geosciences. Under geosciences, the scientific progress achieved by studies using SHEMAT-Suite is discussed; under computational geosciences, the impact of SHEMAT-Suite on testing and developing new algorithms. A summary of important code developments that have not been part of this thesis of SHEMAT-Suite can be found in Table 3.2.

Table 3.2: Important code developments of SHEMAT-Suite outside the work on this thesis. (x) Functionalities not available in the open-source package. (*) Simplified functionality available in the open-source package. (**) SHEMAT-Suite functionality available open-source, additional software required.

Functionality	**Implementation by**
Inverse parameter estimation based on automatic differentiation	Rath et al. (2006)
Latent heat effects due to freezing and melting	Mottaghy and Rath (2006)
Monte Carlo techniques for uncertainty quantification and reduction	Vogt et al. (2010)
Borehole heat exchanger module(*)	Mottaghy and Dijkshoorn (2012)
Shared-memory parallelization	Wolf (2011)
Data assimilation based on the ensemble Kalman Filter	Vogt et al. (2012)
Multi-phase flow module using automatic differentiation(x)	Büsing et al. (2014)
Distributed-memory parallelization(x)	Rostami and Bücker (2014)
Heat transfer model for plane thermo-active geotechnical systems(x)	Kürten et al. (2014, 2015)
Anisotropic flow module using the full permeability tensor(x)	Chen et al. (2016)
Supercritical water/steam module using automatic differentiation(x)	Büsing et al. (2016)
Optimal borehole positioning with respect to reservoir characterization via optimal experimental design(**)	Seidler et al. (2016)
Halite precipitation model in porous sedimentary rock adjacent to salt diapirs(x)	Li et al. (2017)
Efficient two-phase flow in heterogeneous porous media using exact Jacobians(x)	Büsing (2020)

SHEMAT-Suite has contributed to multiple geoscientific areas, the most significant contribution

being in the field of geothermal energy that served as original motivation for the development of the software. To name a few recent examples, temperature sensors installed at borehole heat exchangers have been used to estimate groundwater velocities using SHEMAT-Suite (Michalski and Klitzsch, 2018, 2019). In models of the vadose zone, the heat output of high-voltage electric cables was modeled with SHEMAT-Suite (Hruška et al., 2018, 2019). For deep geothermal reservoirs, the influence of prominent seismic reflectors on the temperature distribution and the formation of convection cells were studied using the code (Ebigbo et al., 2016, Niederau et al., 2017). Additionally, a current research project uses SHEMAT-Suite for predicting the geothermal potential of a caldera system in Mexico (Deb et al., 2019).

Besides geothermics, simulations using SHEMAT-Suite contributed to other geoscientific fields. A property module for ice was developed (Mottaghy and Rath, 2006) and is used currently for ice-sheet modeling. The simulations aim at investigating and understanding the process of subglacial channel formation, which is not only of geological interest but also relevant in context of long-term safety evaluations for nuclear waste deposition. In a current research project, SHEMAT-Suite is used for simulating the submarine groundwater discharge on the New Jersey continental shelf since the last ice age (Thomas et al., 2019). This study relies on the code's ability to simulate density-driven flow and salt transport and contributes to the understanding of freshwater-saltwater interactions in coastal aquifers. The modern aquifer flow simulations of SHEMAT-Suite are often ensemble-based and rely on high-performance computing (Bruckmann and Clauser, 2020). Next to classical geosciences, SHEMAT-Suite has had an impact on computational geoscience as well, mostly by applying and testing algorithms for large-scale simulations. Three-dimensional inverse parameter estimation was performed with SHEMAT-Suite based on automatic differentiation (Rath et al., 2006). Monte Carlo techniques were applied for uncertainty quantification of expected geothermal energy usage in a geothermal reservoir at The Hague, Netherlands (Vogt et al., 2010). A parallelization scheme that was specifically developed for this purpose was implemented in SHEMAT-Suite (Wolf et al., 2008, Bücker et al., 2009, Wolf, 2011, Rostami and Bücker, 2014, Rostami et al., 2014). Due to the heterogeneous nature of the subsurface which is usually sparsely sampled by direct or indirect measurements, inverse and stochastic approaches are becoming increasingly important in the fields of hydrogeology and geothermal reservoir modeling. The ensemble Kalman filter has been implemented and extensively tested for stochastic permeability estimation (Vogt et al., 2010). Optimal experimental design has been used to find the best position of new boreholes in a geothermal reservoir (Seidler et al., 2016). The open-source availability of the code will facilitate the broader usage of these methods in the geosciences.

Now follows a list of some important developments that are currently not part of the open-source version of SHEMAT-Suite presented here, but that are planned to be merged with the open-source publication in the future. A module for multi-phase flow has been developed using SHEMAT-Suite as starting point (Büsing, 2020). Moreover, an equivalent fracture model approach was implemented, which represents fracture permeability by upscaling based on the mimetic finite difference method (Chen et al., 2016, 2018). Since SHEMAT-Suite is a porous medium model with a continuum model approach, it cannot simulate discrete fracture models

and thus flow or transport through fractures explicitly. However, flow through fractured rocks can be approximated under certain conditions by modeling anisotropic permeability or by applying a stochastic-continuum concept (Bruckmann and Clauser, 2020). A basic approach for simulating unconfined aquifers has been implemented. A module for electrokinetic potential simulations was included, which can be used for studies of self potential originating from fluid flow in porous media (Vogt et al., 2014, Büsing et al., 2017). Another module was implemented in SHEMAT-Suite for chemical fluid-rock interaction (Li et al., 2017). In the current research project "Energy oriented Centre of Excellence" (EoCoE[16], Büsing, 2020), the focus shifts to implementing interfaces of SHEMAT-Suite to modern high-performance software, such as the portable data interface (PDI[17]) for data handling, PETSc (Balay et al., 1997) for parallel solver methods and the parallel data assimilation framework (PDAF, Nerger and Hiller, 2013) for a parallel implementation of the ensemble Kalman filter. Finally, the multi-phase flow implementation uses the software PETSc for becoming MPI-parallel and suitable for distributed-memory CPU clusters.

3.5 Guidelines for scientific workflow

To finish this chapter, a summary is given of the following proposed guidelines for developing scientific software that have been derived from developing SHEMAT-Suite:

- git functionality branches ranging from branches of core functionality (SHEMAT-Suite: `master`) to merged branches of combined functionality (SHEMAT-Suite: `master-all`),
- Intensive testing of the core branch and regularly merging up-to-date core functionality into the functionality branches,
- Modular code architecture for multiple simultaneous developments,
- Documentation tailored to scientists, focusing on underlying equations and scientific applications, complementing the classical software documentation,
- Test models repository accessible to all users.

For the future development of SHEMAT-Suite, these guidelines will have multiple effects. First, new users should be attracted to SHEMAT-Suite by the availability and documentation. Secondly, the scientific developers of SHEMAT-Suite receive a stable, but flexible framework for conducting their research. And thirdly, the existing research using SHEMAT-Suite becomes easier to be reproduced. Finally, the software as a whole should grow in functionality, such as multi-phase flow, and in stability by comparing and benchmarking against analytic results and other software.

The recent open-source publication of the SHEMAT-Suite source code has a number of advantages similar to the advantages of the guidelines: (1) it makes SHEMAT-Suite available to a larger

[16] https://www.eocoe.eu/ [Accessed 2020/02/19]
[17] https://pdi.julien-bigot.fr/0.6.0/ [Accessed: 2020/03/17]

community of code users and developers, (2) it makes the existing research using SHEMAT-Suite reproducible, (3) it sparks new interesting developments and additions to the code, and (4) it initiates the formation of a SHEMAT-Suite user community.

Aber wehe, wehe, wehe!
Wenn ich auf das Ende sehe!!

Max und Moritz, Vorwort
Wilhelm Busch

4

Performance comparison of EnKF variants

Performance comparisons of different variants of the ensemble Kalman filter (EnKF) in the literature, in particular applied to permeability estimation, are oftentimes not significant. This is due to the uncertainty of permeability estimation arising from the random seed choice, especially when small ensembles are used.

In this chapter, the influence of the random seed choice on performance comparisons is analyzed. As a result, thresholds are derived for the necessary number of performance comparisons. The chapter is organized as follows. First, in Section 4.1, an introduction to performance comparisons using the EnKF is given. In Section 4.2.1, the EnKF methods that are compared in this chapter are introduced. Section 4.2.2 presents the simulation setups. Subsequently, the tools for comparison of EnKF methods are explained. The results of the various sets of synthetic experiments for the simulation setups, EnKF variants, and ensemble sizes are presented in Section 4.3.

Contents

adapted from | Keller, J., Hendricks Franssen, H.-J., Marquart, G., Comparing seven variants of the ensemble Kalman filter: How many synthetic experiments are needed?, Water Resources Research, 54(9), 6299–6318 (2018).

4.1 Introduction

In this chapter, a large performance comparison of seven EnKF methods is evaluated. Since the EnKF is an ensemble-based method, the ensemble size is an important input of the filter. I recall typical ensemble sizes used in the EnKF literature. Early applications of the EnKF in the atmospheric sciences (Houtekamer and Mitchell, 1998, Hamill and Snyder, 2000, Anderson, 2001, Kalnay, 2002) use fairly large ensemble sizes of up to 500. However, it is typical for large models of the atmospheric or of the geosciences that model complexity restricts ensemble sizes to less than 100. In the field of land surface modeling, where the EnKF is mostly used for state vector estimation, ensemble sizes smaller than 100 are common (e.g. Reichle et al., 2002, De Lannoy and Reichle, 2016). If one considers parameter estimation studies in surface and subsurface hydrology, in many cases ensemble sizes smaller than 250 are used (e.g. Lorentzen et al., 2003, Naevdal et al., 2005, Vrugt et al., 2005, Chen and Zhang, 2006, Shi et al., 2014, Baatz et al., 2017). More uncommon are ensemble sizes larger than 500 (e.g. Devegowda et al., 2010, Vogt et al., 2012).

When small ensemble sizes are used, the EnKF exhibits large fluctuations in the sampled model covariances. These fluctuations might result in filter inbreeding and filter divergence. Filter inbreeding occurs when the spread of the updated ensemble becomes so small that it does not properly represent the error of the EnKF updates. Filter divergence means that the updates of the ensemble center around a wrong parameter value, even though the spread of the ensemble minimizes as if the filter would be approaching the correct parameter value. Different variants of the EnKF have been proposed to reduce problems linked to small ensemble sizes. In this chapter, I compare the following methods: damping of the EnKF (Hendricks Franssen and Kinzelbach, 2008), the local EnKF (Hamill et al., 2001), the hybrid EnKF (Hamill and Snyder, 2000), the dual EnKF (Moradkhani et al., 2005, El Gharamti et al., 2013), the iterative EnKF (Sakov et al., 2012) and the normal score EnKF (Zhou et al., 2011, Schöniger et al., 2012, Li et al., 2012). Implementation details of these algorithms are given in Section 4.2.1. Of course, many methods exist that are not compared in this study. Two examples are modern localization approaches (Chen and Oliver, 2010, Emerick and Reynolds, 2011) and methods

using decompositions of probability distributions into sums of Gaussian distributions (Sun et al., 2009b, Liu et al., 2016).

In order to assess the benefits of new EnKF variants or to compare known variants applied to different geoscience models, it is important to have meaningful comparison methods. Ideally, I would like to compare the inverse solution with the exact, correct inverse solution according to Bayes theorem. However, an analytical solution is not possible for realistic problems and a numerical solution by Markov Chain Monte Carlo methods is extremely expensive. Therefore, for large inverse problems the correct inverse solution will be unknown. Three typical remedies, which can be found in the literature, include comparing estimation results with a synthetic reference, comparing results of a target variable with measurements (e.g., oil production in reservoir engineering), or comparing small-ensemble runs with runs using a larger ensemble. In all cases, EnKF variants are usually compared by evaluating a small number of synthetic experiments. A synthetic experiment is defined here as the comparison of multiple data assimilation methods by applying them to the same parameter estimation problem with fixed synthetic reference field, fixed synthetic observations and fixed ensemble size.

The central motivation of this comparison study is to characterize the random component inherent in the comparison of the performance of different EnKF variants. Such comparisons are often based on small- or medium-sized ensembles and one or a few synthetic experiments. The random component results from the limited ensemble of initial parameter fields, as well as from measurement errors and the associated sampling fluctuations. Applying the EnKF variants to large models, one is computationally restricted to small ensemble sizes and few synthetic experiments, which is why often final results are still subject to considerable uncertainty and a function of specific filter settings. For small models, such as the ones presented in this chapter, it may be possible to suppress the most notorious problems of undersampling by sufficiently enlarging the ensemble size. Nevertheless, knowing how EnKF methods perform on small models with small ensemble sizes might help in deciding which EnKF method to apply to a large model with similar features. However, it cannot be guaranteed that the same EnKF methods that work well on small models would also work well on larger models.

Increasing computer power provides the opportunity to compute many synthetic experiments. In this study, I monitor EnKF performance by running large numbers of synthetic experiments for small and medium ensemble sizes. I hypothesize that randomness is non-negligible in the evaluation of a small number of synthetic experiments. The influence of the random component is characterized by calculating RMSEs between estimated permeability fields and the synthetic true field for each synthetic experiment. For 1, 10 and 100 synthetic experiments, it is evaluated how strongly means of these smaller numbers of RMSEs deviate from the mean RMSE over all 1 000 synthetic experiments. Finally, I determine the performance difference between EnKF variants (in terms of RMSE), which can be detected with a given number of synthetic experiments (1, 10, or 100).

4.2 Methods

4.2.1 EnKF variants

The following subsections contain introductions to the variants of the EnKF algorithm compared in this case study. For a general introduction to the EnKF, see Section 2.2.

Damping

A damping factor $0 < \alpha \leq 1$ can be included in the EnKF to counteract filter divergence (Hendricks Franssen and Kinzelbach, 2008). The damped assimilation step is given by

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \alpha \mathbf{K} \left(\mathbf{d}_i - \mathbf{H}\mathbf{x}_i^f \right). \tag{4.1}$$

In this study, $\alpha \in (0, 1]$ only dampens the parameter updates - the state updates are kept undamped. For groundwater flow simulation, damping of the parameter update reduces the impact of ensemble-based linearization between the hydraulic head and hydraulic conductivity. The ensemble covariance matrix necessarily treats any relation between two states or a state and a parameter as linear, but for flow in heterogeneous media this relation is non-linear. Updating the parameters by smaller steps, more slowly approximating the posterior values is therefore expected to be more stable. In synthetic experiments it is commonly found that this also reduces filter inbreeding and filter divergence (Hendricks Franssen and Kinzelbach, 2008, Wu and Margulis, 2011).

Localization

For small ensemble sizes, undersampling can lead to large fluctuations of ensemble covariances. Even for locations that are far apart in space, nonzero covariances may appear, but such covariances, or correlations, are most likely spurious (Houtekamer and Mitchell, 1998). Localization methods reduce the effect of these spurious long-range correlations on the filter update. To this end, a correlation matrix $\rho \in \mathbb{R}^{n_s \times n_m}$ (Gaspari and Cohn, 1999) is multiplied elementwise with the first part of the Kalman gain (Hamill et al., 2001):

$$\mathbf{K}_{loc} = \left[\rho \circ \left(\mathbf{P}_e \mathbf{H}^T \right) \right] \left(\mathbf{H}\mathbf{P}_e\mathbf{H}^T + \mathbf{R} \right)^{-1}. \tag{4.2}$$

Typically, a characteristic length scale $\lambda \in \mathbb{R}^+ \, [L]$ is associated with the correlation matrix. For this study, both ρ and its length scale are chosen as in Gaspari and Cohn (1999). An entry of the correlation matrix ρ is a function of the distance $d \in \mathbb{R}^+ \, [L]$ between two locations and the

correlation length scale λ. The formula for an entry of ρ is given by

$$\rho(d, a = \sqrt{\frac{10}{3}}\lambda) = \begin{cases} -\frac{\left(\frac{d}{a}\right)^5}{4} + \frac{\left(\frac{d}{a}\right)^4}{2} + \frac{5\left(\frac{d}{a}\right)^3}{8} + \frac{5\left(\frac{d}{a}\right)^2}{3} + 1, & 0 \leq \frac{d}{a} < 1 \\ \frac{\left(\frac{d}{a}\right)^5}{12} + \frac{\left(\frac{d}{a}\right)^4}{2} + \frac{5\left(\frac{d}{a}\right)^3}{8} + \frac{5\left(\frac{d}{a}\right)^2}{3} - 5\frac{d}{a} + 4 - \frac{2}{3\frac{d}{a}}, & 1 \leq \frac{d}{a} < 2 \\ 0, & 2 \leq \frac{d}{a}. \end{cases} \tag{4.3}$$

The parameter a is a multiple of λ adapted to the functional form of ρ. The objective of ρ is to approximate a two-dimensional Gaussian bell curve:

$$G(d, \lambda) = \exp\left(-\frac{d^2}{2\lambda^2}\right). \tag{4.4}$$

Contrary to the approximated Gaussian correlations, all values of ρ are zero for distances greater than $2a$.

Hybrid EnKF

For the hybrid EnKF (Hamill and Snyder, 2000), the covariance matrix is chosen as a sum of the usual ensemble covariance matrix and a static background covariance matrix:

$$\mathbf{P}_{\text{hybrid}} = \beta\mathbf{P}_e + (1 - \beta)\mathbf{P}_{\text{static}}. \tag{4.5}$$

The factor $0 \leq \beta \leq 1$ determines the weight assigned to the ensemble covariance matrix. The static background covariance matrix represents prior knowledge about the geology and physics of the model.

Dual EnKF

Another method entering the comparison is the dual EnKF by Moradkhani et al. (2005) and Wan and Nelson (2001). The state vector $\mathbf{x}_i^f \in \mathbb{R}^{n_s}$ is split into two parts: $\mathbf{x}_{s,i}^f \in \mathbb{R}^{n_{s,s}}$ contains the state variables, and $\mathbf{x}_{p,i}^f \in \mathbb{R}^{n_{s,p}}$ contains the parameters. Here, $n_{s,s} \in \mathbb{N}$ is the number of dynamic variables in the state vector, $n_{s,p} \in \mathbb{N}$ is the number of parameters in the state vector, and $n_{s,s} + n_{s,p} = n_s$. When the filter reaches a measurement time, only the parameters are updated according to

$$\mathbf{x}_{s,i}^f \qquad \mathbf{x}_{p,i}^a = \mathbf{x}_{p,i}^f + \mathbf{K}_p\left(\mathbf{d}_i - \mathbf{H}_p\mathbf{x}_{p,i}^f\right). \tag{4.6}$$

$\mathbf{K}_p \in \mathbb{R}^{n_{s,p} \times n_m}$ and $\mathbf{H}_p \in \mathbb{R}^{n_m \times n_{s,p}}$ are the parts of the Kalman gain and measurement matrix projected onto the parameter space. After the parameter update, the previous forward simulation is rerun using the same dynamic states, but the updated parameters. In the second and final updating step, only the states $\mathbf{x}_{s,i}^{f,2}$ are updated according to

$$\mathbf{x}_{s,i}^a = \mathbf{x}_{s,i}^{f,2} + \mathbf{K}_s\left(\mathbf{d}_i - \mathbf{H}_s\mathbf{x}_{s,i}^{f,2}\right) \qquad \mathbf{x}_{p,i}^a. \tag{4.7}$$

The matrices $\mathbf{K}_s \in \mathbb{R}^{n_{s,s} \times n_m}$ and $\mathbf{H}_s \in \mathbb{R}^{n_m \times n_{s,s}}$ are projected onto the space of state variables. Note that the same measurement values are used for both the parameter and the state variable update. The whole procedure is repeated for each assimilation time.

Normal score EnKF

The normal score EnKF (NS-EnKF, Zhou et al., 2011) was developed to handle non-Gaussian probability distributions inside an EnKF framework. The method inherits its name from the normal score transform (Journel and Huijbregts, 1978, Goovaerts, 1997, Deutsch and Journel, 1995). The ensemble of states, parameters, and measurement values are transformed before assimilation starts. The transform uses the cumulative distribution functions to turn the ensemble into one of a normalized Gaussian pdf.

$$NS(x_i^f) := G^{-1}(F(x_i^f)). \tag{4.8}$$

Here, $x_i^f \in \mathbb{R}$ is a single component of the ith realization of the state and parameter vector $\mathbf{x}_i^f$. F is the empirical cumulative distribution function of x_i^f, and G is the cumulative distribution function of the Gaussian pdf with zero mean and unit standard deviation. The normal score transform largely preserves the correlation structure of the variables by keeping intact the relative order of the ensemble members. The EnKF update is carried out in terms of the transformed values. After the update, states and parameters are back-transformed:

$$NS^{-1}(x_i^a) := F^{-1}(G(x_i^a)). \tag{4.9}$$

While the back-transform is similar to the transform, a complication arises, since F is only an ensemble approximation (step-function) of a cumulative distribution function. If $G(x_i^a)$ falls into the range of the ensemble, F^{-1} is interpolated between the values of the two closest ensemble members. When updated ensemble members are located outside the range of the ensemble, an extrapolation method is needed for performing the back-transform. To illustrate the extrapolation method, I suppose without loss of generality that I start with a sorted the ensemble x_i^f according to the index i. Then, the empirical cumulative distribution F can be written as

$$F(x_i^f) = \frac{i - 0.5}{n_e} \tag{4.10}$$

and I can see that updated values x_i^a fall outside the interpolation range if either

$$G(x_i^a) < \frac{0.5}{n_e} = F(x_0^f) = \min_i F(x_i^f) \quad \text{or} \quad G(x_i^a) > \frac{n_e - 0.5}{n_e} = F(x_{n_e}^f) = \max_i F(x_i^f). \tag{4.11}$$

In these cases, the following extrapolation method is used to calculate F^{-1} for outliers. Two artificial support points $x_0^s \in \mathbb{R}$ and $x_1^s \in \mathbb{R}$ are selected and define artificial boundary values of the cumulative function F:

$$F(x_0^s) = 0 \qquad F(x_1^s) = 1. \tag{4.12}$$

The distance between an artificial support point and the neighboring x_i^f is set to three times the spread of the original support points of the range of F^{-1}, i.e. the x_i^f. In this way, a cut-off applies and the back-transform of the x_i^a can never have a value outside of the interval between the artificial support points:

$$NS^{-1}(x_i^a) \in [x_0^s, x_1^s]\,. \tag{4.13}$$

By choosing the support points as far apart as explained above, this does not overly restrict the updates. Once the distributions are properly back-transformed, the computation of the forward model continues.

Iterative EnKF

The iterative version of the EnKF (IEnKF) is inspired by the exposition in Sakov et al. (2012). Iteration is introduced to the EnKF to mitigate problems related to the non-linearity of model equations. In the version from this work, after an update of parameters by the EnKF, the simulation restarts from the beginning using the newly updated parameters as input (in contrast, the dual EnKF always restarts from the previous update). One drawback of this method is its computational demand. Let $T \in \mathbb{R}^+$ $[T]$ be the computing time of a normal EnKF run. Then, I can derive estimates for the computing times of the dual and the iterative EnKF. For the dual EnKF, every update and forward computation is carried out twice. Thus, I expect a computing time of $T_{\text{Dual}} = 2T$. Assuming equidistant assimilation times for iterative EnKF, the computing time depends on the number of assimilation times n_t: $T_{\text{Iterative}} = T \cdot (\frac{1}{n_t} + \frac{2}{n_t} + \cdots + 1) = T \cdot \frac{n_t+1}{2}$. For the synthetic study in this chapter ($n_t = 100$), the ratio between computation times amounts to $\frac{T_{\text{Iterative}}}{T_{\text{Dual}}} = \frac{n_t+1}{4} \eqsim 25$.

4.2.2 Design of the synthetic experiments

This section deals with the two setups for synthetic experiments computed in this study, a forward model combining groundwater flow and solute transport and a groundwater flow model with an injection well.

Tracer model

The 2-D subsurface model for groundwater flow and solute transport is based on a grid consisting of 31 × 31 cells of size 2 m × 2 m. These 961 cells define a square simulation domain of size 62 m × 62 m. The governing equations are solved for a simulation time of 1 200 days divided into 200 equal intervals.

Boundary and initial conditions include a prescribed hydraulic head of 11 m at the southern boundary and 10 m at the northern boundary, creating a flow from south to north. The two remaining boundaries are impermeable. A tracer concentration of 80×10^{-3} mol/L is prescribed on the southern boundary; at the northern boundary I set the concentration to 60×10^{-3} mol/L.

The last value also serves as initial concentration throughout the model domain leading to solute transport from south to north.

The tracer is subjected to slow diffusion with a diffusion coefficient of $1.5 \times 10^{-9}\,\mathrm{m}^2/\mathrm{s}$. The small diffusion coefficient is the reason that the temporal evolution of the tracer concentration is almost completely determined by advection. The fluid is water with the standard properties set in SHEMAT-Suite. The porosity of the rock is 10 %. At two locations with coordinates $(19\,\mathrm{m}, 31\,\mathrm{m})$ and $(43\,\mathrm{m}, 31\,\mathrm{m})$, the tracer concentration is recorded every 12 days, summing up to 100 measurement times in total.

Well model

The 2-D groundwater flow well model is based on a grid consisting of 31×31 cells of size $20\,\mathrm{m} \times 20\,\mathrm{m}$ resulting in a model domain of $620\,\mathrm{m} \times 620\,\mathrm{m}$. The governing equations are solved for a simulation time of 18 days divided into 1 200 equal intervals. Boundary and initial conditions include a prescribed hydraulic head of 11 m at a central well at coordinates $(310\,\mathrm{m}, 310\,\mathrm{m})$. All remaining boundary conditions and initial conditions are set to a hydraulic head of 10 m. Again, the fluid is water with its standard properties and the porosity of the rock is 10 %. In this model, there are 48 measurement locations located along a regular 7×7 grid excluding the central well position. Hydraulic heads are recorded at 60 measurement times every 7 hours and 12 minutes.

The numerical settings for the solution of the linear system of equations are identical for both the groundwater flow and species transport equation in the two cases. In the Picard iteration, I demand a relative tolerance of 1×10^{-10}. The linear solution method Bi-CGStab is called with two termination criteria: An accepted relative error of 1×10^{-14} and a maximum number of 500 iterations.

4.2.3 Simulation of permeability fields

The heterogeneous synthetic reference permeability fields, displayed in Figure 4.1, are generated by sequential multi-Gaussian simulation using the software SGSIM (Deutsch and Journel, 1995). The following spherical correlation function is used:

$$\rho(d) = \begin{cases} 1 - \left(\frac{d}{a}\right) \cdot \left[\frac{3}{2} - \frac{1}{2} \cdot \left(\frac{d}{a}\right)^2\right]. & 0 \leq \frac{d}{a} < 1 \\ 0, & 1 \leq \frac{d}{a}. \end{cases} \tag{4.14}$$

Here $d \in \mathbb{R}^+$ denotes the Euclidean distance between two grid cells and $a \in \mathbb{R}^+$ is the correlation length of the permeability field. No nugget effect is considered for the generation of the permeability field. Both permeability fields are generated using a mean of $-12.0 = \log_{10}(K/\mathrm{m}^2)$ (without unit) and standard deviation 0.5. Here and in the following, a logarithmic permeability without unit like $-12.0 = \log_{10}(K/\mathrm{m}^2)$ corresponds to the actual permeability $10^{-12}\,\mathrm{m}^2$. Correlation lengths of $a = 50\,\mathrm{m}$ for the tracer model and $a = 60\,\mathrm{m}$ for the well model are used.

The correlation length of the tracer model was chosen such that relatively large areas of distinct permeability would be generated. Theses areas either block the tracer or constitute flow paths. Note that, relative to the full grid size, the correlation length of the well model is smaller by 1 order of magnitude. This was chosen for showing how the EnKF methods would react to a more heterogeneous permeability field relative to the full model size.

The ensembles of initial permeability fields for the 1 000 synthetic experiments (for each physical problem setup, each ensemble size, and each EnKF variant), which I use to compare performance, are generated by SGSim using one set of input parameters but varying random seeds. A mean logarithmic permeability of $-12.5 = \log_{10}(K/\mathrm{m}^2)$ (compared to the reference fields with $-12.0 = \log_{10}(K/\mathrm{m}^2)$) and a standard deviation of 0.5 (identical to the reference fields) are used. Consequently, the EnKF algorithms have to adjust the overall mean of the permeability fields as well as the differences that would have been present even for equal mean values due to spatial variability.

EnKF setup

The EnKF state vector includes permeability, hydraulic head, and, for the tracer model, concentration. Permeability is the target parameter of the estimation. Concentration or hydraulic head values serve as the observations that drive the estimation. Measurement noises are assumed constant: $\sigma_{\mathtt{c}} = 7.1 \times 10^{-3}$mol/L for concentration and $\sigma_{\mathtt{h}} = 5 \times 10^{-2}$ m for hydraulic head. These noises resemble the uncertainty of typical measurement devices.

Most EnKF methods require the specification of additional parameters. The damping factor of the damped EnKF is set to $\alpha = 0.1$ (cf. Hendricks Franssen and Kinzelbach, 2008). In the local EnKF, the length scale λ is set to 150 m for the well model and the tracer model. For the hybrid EnKF, a constant diagonal background covariance matrix is used. The value of β used in this study is 0.5. Results of synthetic experiments, for which λ and β are varied, can be found in Section 4.3.5.

In this study a series of EnKF data assimilation experiments, called synthetic experiments, is performed. For each of the methods presented in Section 4.2.1, seven ensemble sizes (50, 70, 100, 250, 500, 1 000 and 2 000 realizations) are tested. For the four smaller ensemble sizes, 1 000 synthetic experiments are carried out. For the larger ensemble sizes, due to computational limitations, only 100 synthetic experiments are computed.

4.2.4 Performance evaluation

In a single synthetic experiment called j with given ensemble size, the RMSE for one of the EnKF methods in the synthetic experiment is given by

$$\mathrm{RMSE}_j = \sqrt{\frac{1}{n_g} \sum_{i=1}^{n_g} \left(\bar{Y}_{ij} - Y_{ij}^t \right)^2}. \tag{4.15}$$

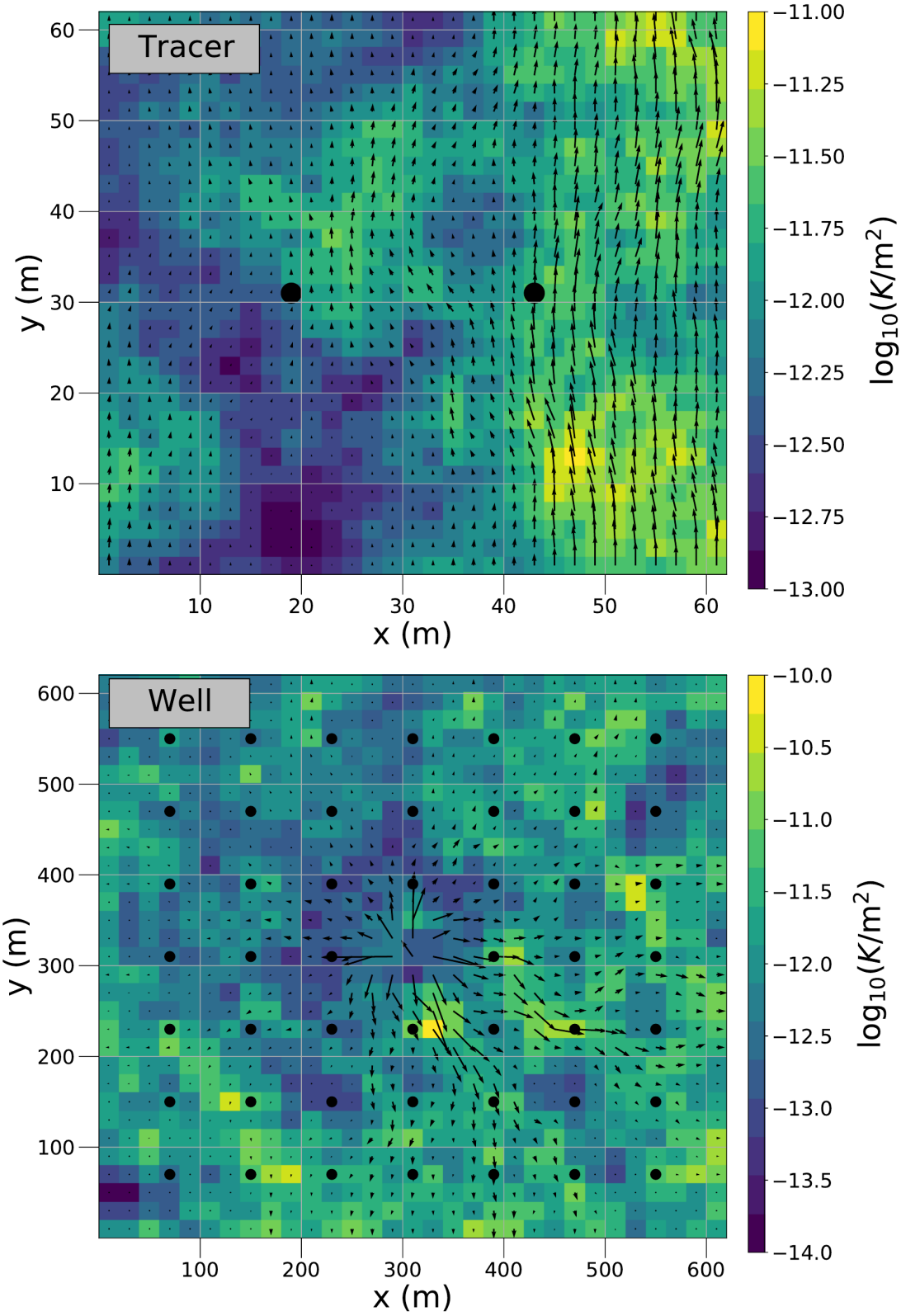

Figure 4.1: The logarithmic reference permeability fields for the tracer model and the well model with groundwater flow vectors. Measurement locations are depicted as black circles.

Here $n_g \in \mathbb{N}$ is the number of grid cells, $\bar{\mathbf{Y}}_j \in \mathbb{R}^{n_g}$ is the vector containing the estimated mean logarithmic permeabilities across the model domain and $\mathbf{Y}_j^t \in \mathbb{R}^{n_g}$ is the vector containing the corresponding synthetic reference logarithmic permeabilities.

For a given ensemble size, a large number of synthetic experiments is computed. These synthetic experiments differ solely in the perturbation of initial fields and measurements. A single synthetic experiment j provides a sample-RMSE for each EnKF method, calculated according to Equation (4.15) and called $r_j^{a,\,n_e} \in \mathbb{R}^+$. Taken together, the $n_{syn} = 1\,000$ synthetic experiments (or in the case of large ensembles $n_{syn} = 100$ synthetic experiments) for a given ensemble size n_e provide an approximate probability density function of the RMSE for each method a. I calculate RMSE means according to

$$\bar{r}^{a,\,n_e} = \frac{1}{n_{syn}} \sum_{j=1}^{n_{syn}} r_j^{a,\,n_e}. \tag{4.16}$$

These RMSE means are used to compare EnKF methods.

The question arises, whether a small number of synthetic experiments n_{syn} would suffice to evaluate the performance of the EnKF methods. I compare RMSE means of two EnKF methods on the basis of $n_{syn} = 1$, 10, or 100 synthetic experiments. Ten thousand subsets X of n_{syn} synthetic experiments are randomly sampled from the 1 000 synthetic experiments. I calculate and compare the corresponding means $\bar{r}_X^{a,\,n_e} \in \mathbb{R}^+$. The fraction, for which one EnKF method a yields a smaller RMSE than another EnKF method b

$$p^{a<b,\,n_e} = \frac{\#\{X \mid \bar{r}_X^{a,\,n_e} < \bar{r}_X^{b,\,n_e}\}}{10\,000}, \tag{4.17}$$

is recorded (the sum $p^{a<b,\,n_e} + p^{b<a,\,n_e}$ is by definition equal to 1.0). In doing so, I estimate the probability that one method outperforms another method on the basis of n_{syn} synthetic experiments.

Quotients of the RMSE means (based on all 1 000 synthetic experiments) for all combinations of EnKF variants are calculated:

$$q^{a<b,\,n_e} = \frac{\bar{r}^{a,\,n_e}}{\bar{r}^{b,\,n_e}}. \tag{4.18}$$

In this formula I choose EnKF method a to have the smaller RMSE mean, so that $q^{a<b,\,n_e} \leq 1$. From the quotients, I calculate relative differences:

$$d^{a<b,\,n_e} = \frac{\bar{r}^{b,\,n_e} - \bar{r}^{a,\,n_e}}{\bar{r}^{b,\,n_e}} = 1 - q^{a<b,\,n_e}. \tag{4.19}$$

For example, it could be that EnKF variant a gives on average (calculated over 1 000 synthetic experiments) a RMSE that is $d^{a<b,\,n_e} = 10\,\%$ smaller than another EnKF variant b. In that case, the quotient would be $q^{a<b,\,n_e} = 0.9$. I analyze which relative differences across pairs of EnKF methods a and b are with at least 95 % probability statistically significant for n_{syn} synthetic experiments and which are not.

$$D_+^{n_e} = \{d^{a<b,\,n_e} \mid p^{a<b,\,n_e} > 0.95\}, \tag{4.20}$$

$$D_{-}^{n_e} = \{d^{a<b,\,n_e} \mid p^{a<b,\,n_e} \leq 0.95\}. \tag{4.21}$$

Finally, I compare the smallest significant relative difference

$$d_{+}^{min} = \min D_{+}^{n_e} \tag{4.22}$$

to the largest insignificant relative difference

$$d_{-}^{max} = \max D_{-}^{n_e}. \tag{4.23}$$

In general, one would expect d_{+}^{min} to be larger than d_{-}^{max}. In this case, I choose the threshold for significant relative differences between the two values. However, due to the specific shape of the RMSE distributions (especially related to their spread), it may occur that $d_{+}^{min} < d_{-}^{max}$. If this happens, I evaluate the comparisons on the basis of their specific distributions and choose the threshold manually.

4.3 Results and discussion

4.3.1 Tracer model

Comparison of EnKF variants in terms of mean RMSE

Figure 4.2 shows the mean RMSEs for all methods and ensemble sizes 50, 70, 100 and 250. The RMSE values $\bar{r}^{a,\,n_e}$ range from 0.3 to 0.4. For ensemble size 50, the hybrid EnKF yields the smallest mean RMSE (0.328) followed by the iterative EnKF, the damped EnKF and the normal score EnKF (0.351, 0.351) and 0.356). Then, the local EnKF (0.363) is followed by the classical EnKF (0.366). The dual EnKF yields the largest RMSE: 0.388.

When the ensemble size is increased from 50 to 250, all methods produce ensemble mean permeability fields closer to the reference. The RMSE for the damped EnKF is by $\bar{r}^{\mathrm{Damped},\,50} - \bar{r}^{\mathrm{Damped},\,250} = 0.016$ smaller for an ensemble size of 250 than for an ensemble size of 50. This RMSE reduction is the smallest of all EnKF methods and the hybrid EnKF shows the next smallest reduction ($\bar{r}^{\mathrm{Hyb-EnKF},\,50} - \bar{r}^{\mathrm{Hyb-EnKF},\,250} = 0.029$ smaller for ensemble size 250 than for ensemble size 50). As a result, the damped EnKF has the second largest RMSE among all EnKF methods for ensemble size 250. The ranking of the remaining methods is identical for all four ensemble sizes.

The second part of Figure 4.2 contains tables of quotients $q^{a<b,\,50}$ and $q^{a<b,\,250}$ (as defined in Equation (4.18)) for all pairs of EnKF methods. The largest relative difference in the tables is 16 % (*hybrid EnKF versus dual EnKF*) for both ensemble sizes 50 and 250. Three relative differences are as small as 1 % (*iterative EnKF versus damped EnKF* and *local EnKF versus classical EnKF* for ensemble size 50, *local EnKF versus classical EnKF* for ensemble size 250).

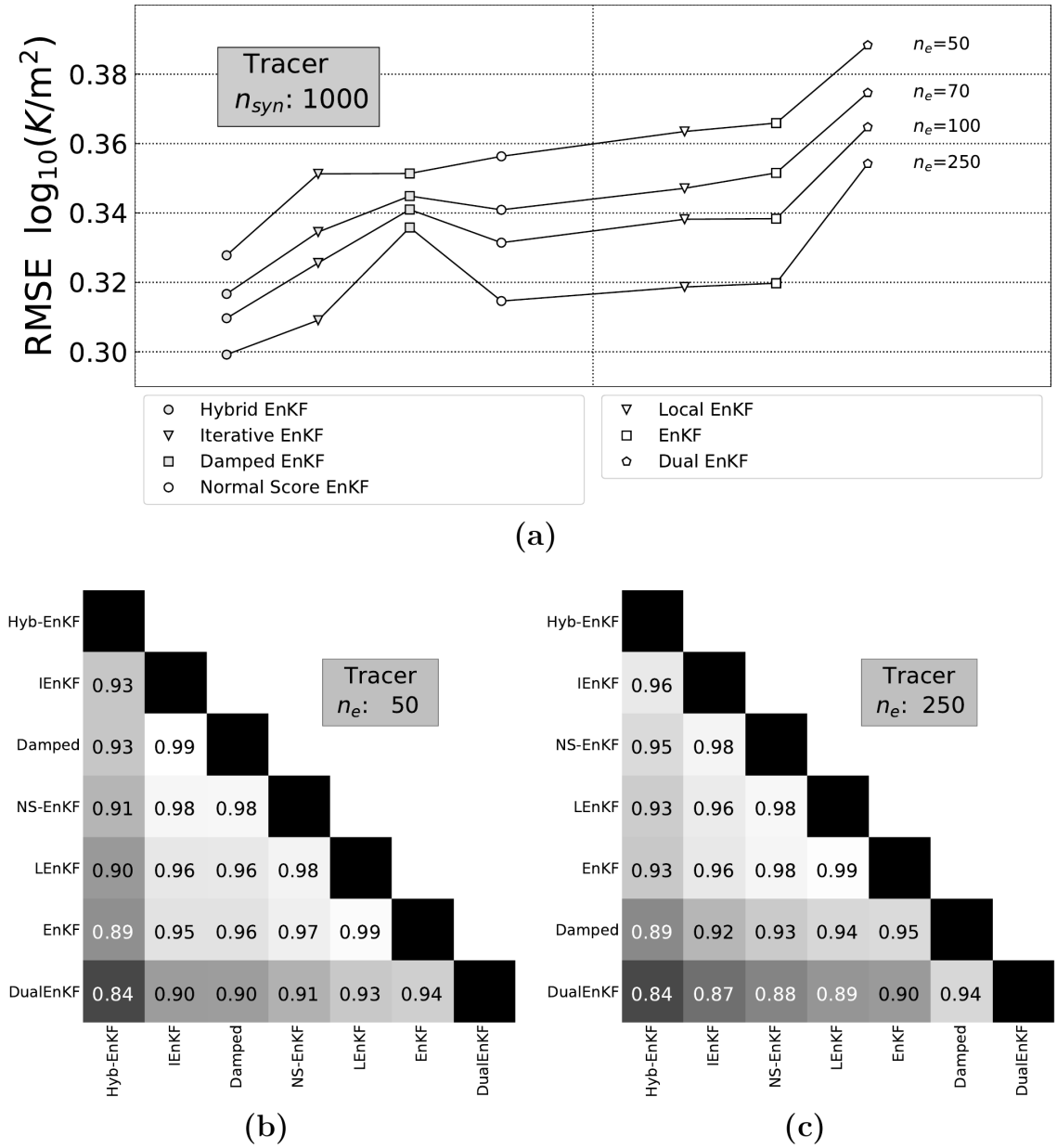

Figure 4.2: Mean root-mean-square errors (RMSEs) for seven ensemble Kalman filter (EnKF) methods derived from 1 000 synthetic experiments for the tracer model - ensemble sizes: $n_e = 50$, 70, 100, and 250 (a). For ensemble sizes $n_e = 50$ (b) and 250 (c), the graph shows the RMSE mean for a method on the horizontal axis divided by the RMSE mean for a method on the vertical axis.

In the following, I examine which relative differences can be considered significant as function of the number of synthetic experiments.

It is worth mentioning that the mean RMSE difference between initial permeability fields and the synthetic truth (0.62) is significantly reduced in all synthetic experiments. Finally, I note that the large RMSEs of the local EnKF, the classical EnKF, and the dual EnKF can be attributed at least partially to a non-negligible number of synthetic experiments where the methods yield very large RMSEs. The full RMSE distributions are shown in Figure 4.7.

Thresholds on significant RMSE differences

It is evaluated whether significant differences in performance between EnKF variants can be demonstrated for 1, 10, or 100 synthetic experiments. Figure 4.3 shows probabilities to find a smaller RMSE mean for a given EnKF variant compared to another EnKF variant based on n_{syn} synthetic experiments. This output has to be understood in the sense of Equation (4.17). The upper triangle in one of the squares contains the fraction of comparisons for which the EnKF method on the left of the row of the square yielded the smaller mean RMSE, while the lower triangle contains the fraction for which the EnKF method on the bottom of the column of the square yielded the smaller mean RMSE.

The probabilities $p^{a<b,\,50}$ and $p^{a<b,\,250}$ from Figure 4.3 are cross-checked with the quotients $q^{a<b,\,50}$ and $q^{a<b,\,250}$ in Figure 4.2 to determine thresholds for significant relative RMSE differences. Significant relative differences are defined as the relative differences, for which at least 95 % of the comparisons in the squares of Figure 4.3 favor one of the two EnKF methods. Based on a single synthetic experiment, and for ensemble size 50, all relative RMSE differences are insignificant. Consequently, RMSE differences between two EnKF methods smaller than 16 % cannot be detected on the basis of a single synthetic experiment and ensemble size 50. The 16 % are a restatement of the smallest quotient in 0.86 in Figure 4.2. For ensemble size 250, the RMSE difference of 11 % for the comparison pair *hybrid EnKF versus damped EnKF* (compare quotient of 0.89 in Figure 4.2) is significant. The significance is derived from the corresponding square in Figure 4.3 for $n_e = 250$ and $n_{syn} = 1$ that says that 95.5 % of comparisons yield a smaller RMSE for hybrid EnKF. Significance is derived the same way for all statements that follow. On the other hand, there are three insignificant comparisons with larger relative differences (*hybrid EnKF versus dual EnKF*, *iterative EnKF versus dual EnKF* and *normal score EnKF versus dual EnKF*). The anomaly that three insignificant relative differences are larger than a significant relative difference is due to the narrow RMSE distribution of the damped EnKF (Figure 4.7 shows that there are fewer RMSEs below 0.3 for the damped EnKF than for any other method). Taking this into account, I set the threshold for which significant RMSE differences can be detected to 15 %. For 10 synthetic experiments and ensemble size 50, there are RMSE differences of 7 % and 9 % that are significant (*hybrid EnKF versus damped EnKF*, *hybrid EnKF versus normal score EnKF*) and differences of 9 % and 10 % that are not (*normal score EnKF versus dual EnKF*, *iterative EnKF versus dual EnKF*). Taking into account the narrow RMSE distribution of the damped EnKF, the threshold is set to 9 %. For ensemble size 250, and 10 synthetic experiments,

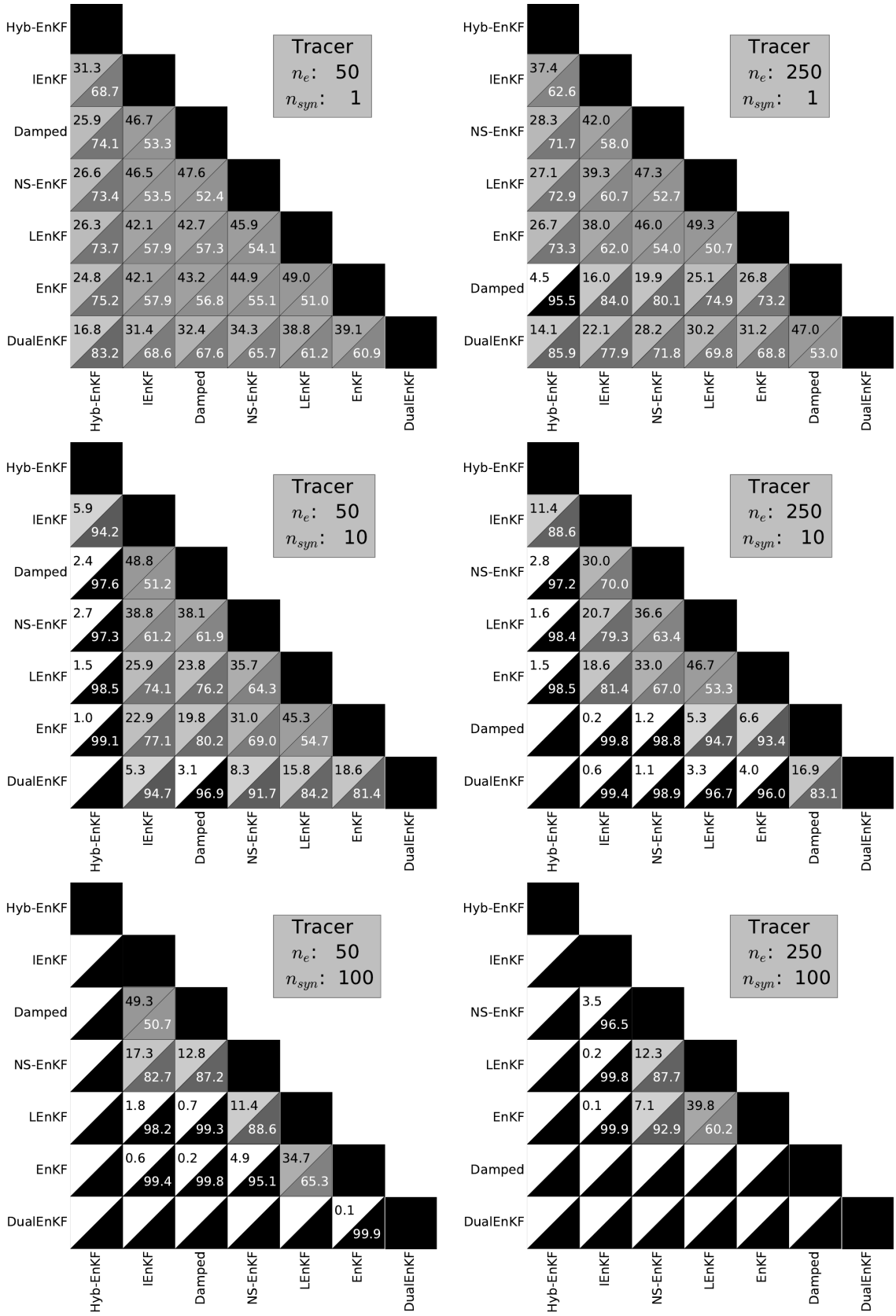

Figure 4.3: Probabilities to find a smaller root-mean-square error (RMSE) mean for a given ensemble Kalman filter (EnKF) variant compared to another EnKF variant, based on 1, 10, and 100 synthetic experiments (tracer model, ensemble sizes 50 and 250). In a square, the probabilities in the upper triangle correspond to the EnKF method that names the row, the ones in the lower triangle correspond to the method that names the column. The gray-scale is a visualization of the probabilities, triangles of significant probabilities larger than 95 % are black.

there is a RMSE difference of 5 % that is significant (*hybrid EnKF versus normal score EnKF*) and there are two of 6 % that are not (*hybrid EnKF versus normal score EnKF*, *damped EnKF versus dual EnKF*). The threshold is set to 6 %. For 100 synthetic experiments and ensemble size 50, there is a difference of 3 % that is significant (*normal score EnKF versus classical EnKF*) and there are three differences of 2 % that are not (*iterative EnKF versus normal score EnKF*, *damped EnKF versus normal score EnKF*, *normal score EnKF versus local EnKF*). The threshold is set to 2 %. For ensemble size 250, there is a difference of 2 % that is significant (*iterative EnKF versus normal score EnKF*) and there are two differences of 2 % that are not (*normal score EnKF versus local EnKF*, *normal score EnKF versus classical EnKF*). The threshold is set to 2 %. All thresholds are provided in Table 4.1.

4.3.2 Well model

Comparison of EnKF variants in terms of mean RMSE

For the well model, EnKF variants (and ensemble sizes) are compared on the basis of the calculated RMSE means in the same way as for the tracer model. Results for ensemble sizes 50, 70, 100, and 250 are discussed here. Results for ensemble sizes 500, 1 000, and 2 000 are discussed later in Section 4.3.3.

Figure 4.4 shows the mean RMSEs for all methods and ensemble sizes. The values range from 0.55 to 0.95. For the smallest ensemble size of 50, observed mean RMSEs show the largest spread. The local EnKF yields the smallest mean RMSE (0.63 followed by the hybrid EnKF, the normal score EnKF and the damped EnKF (0.69, 0.70, and 0.75). The iterative EnKF, the classical EnKF, and the dual EnKF (0.81, 0.84, and 0.92) yield mean RMSEs that are larger than those for the initial permeability fields suggesting divergence of the algorithm for a significant fraction of the synthetic experiments.

When the ensemble size is increased from 50 to 250, all methods get closer to the reference. For the local EnKF, the hybrid EnKF, the normal score EnKF and the damped EnKF the RMSE reduces by up to 0.1, when the ensemble size is increased from 50 to 250. For the iterative EnKF, the classical EnKF and the dual EnKF this reduction is around 0.2. As a result, the iterative EnKF ends up with the third smallest RMSE for ensemble size 250, the classical EnKF performs better than the damped EnKF and the dual EnKF still has the largest RMSE, but much closer to the other methods.

The second part of Figure 4.4 contains tables of quotients $q^{a<b,\,50}$ and $q^{a<b,\,250}$ (as defined in Equation (4.18)) for all pairs of EnKF methods. For the well model, the relative differences are much larger than for the tracer model. The largest relative differences in the tables are 32 % for ensemble size 50 and 19 % for ensemble size 250 (both *local EnKF versus dual EnKF*). For ensemble size 50, the smallest relative difference is 3 % for the comparison pair *hybrid EnKF versus normal score EnKF*. For ensemble size 250, two relative differences are at 2 % (*hybrid*

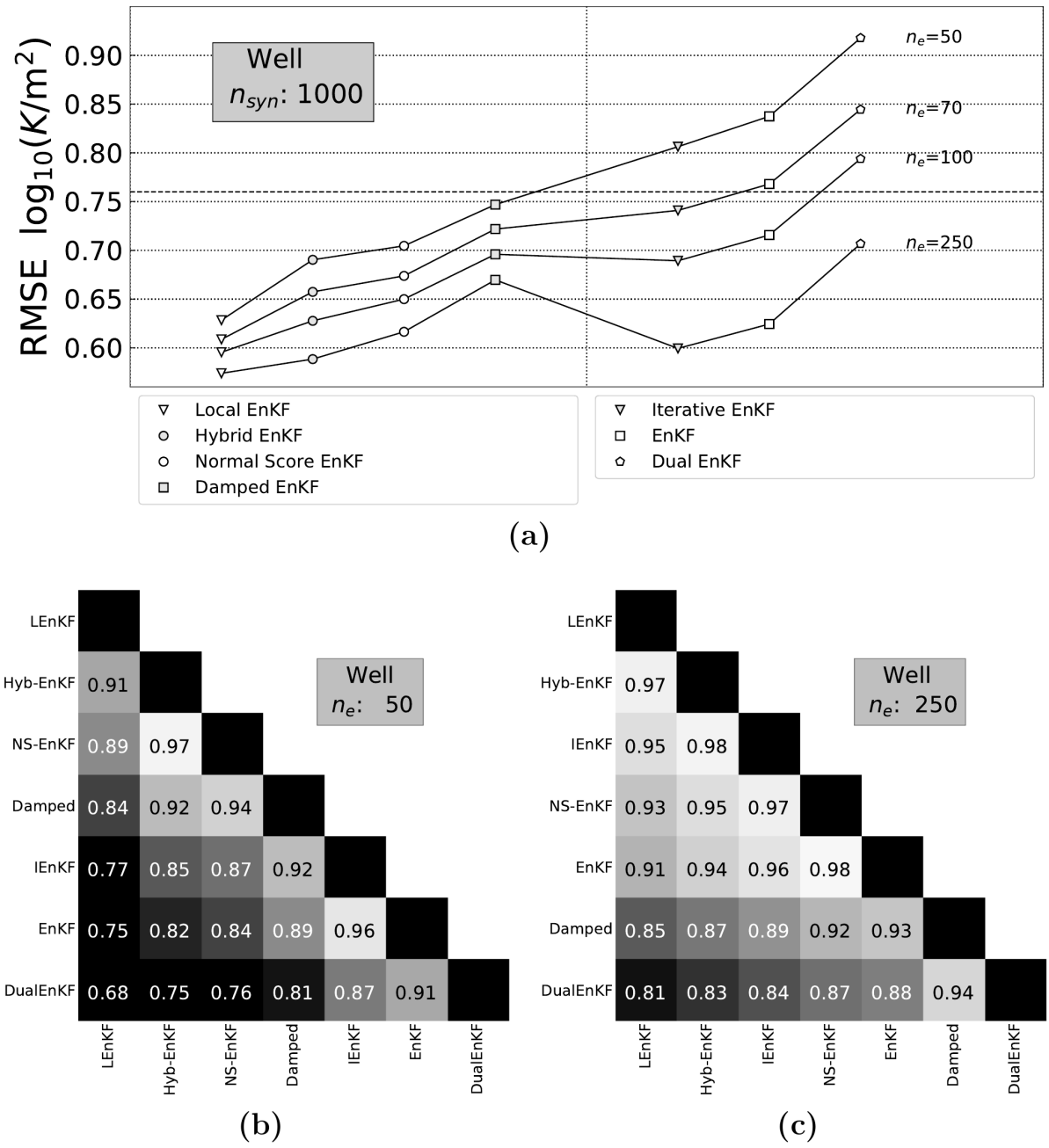

Figure 4.4: Mean root-mean-square errors (RMSEs) for seven ensemble Kalman filter (EnKF) methods derived from 1 000 synthetic experiments for the well model - ensemble sizes: $n_e = 50$, 70, 100, and 250 (a). For ensemble sizes $n_e = 50$, (b), and 250 (c), the graph shows the RMSE mean for a method on the horizontal axis divided by the RMSE mean for a method on the vertical axis.

EnKF versus iterative EnKF, *normal score EnKF versus classical EnKF*). In general, relative differences are smaller for the larger ensemble size.

The much lower RMSE for ensemble size 250 (compared to ensemble size 50) is most probably related to the overall worse performance of the data assimilation experiments for the well model compared to the tracer model. The worse performance is related to the much smaller correlation length, in combination with the high variance of the permeability field, for this setup. The benefit of estimating noisy model covariances with a larger ensemble seems therefore to be higher than for the tracer model.

Thresholds on significant RMSE differences

Also for the well model, it is evaluated whether significant differences in performance between EnKF variants can be demonstrated with 1, 10 or 100 synthetic experiments.

The probabilities $p^{a<b,\,50}$ and $p^{a<b,\,250}$ from Figure 4.5 are cross-checked with the quotients $q^{a<b,\,50}$ and $q^{a<b,\,250}$ in Figure 4.4 to determine thresholds for significant relative RMSE differences. For a single synthetic experiment and ensemble size 50, there are significant RMSE differences of 25 % and 23 % (*local EnKF versus classical EnKF*, *local EnKF versus iterative EnKF*) and insignificant RMSE differences of 24% and 25% (*normal score EnKF versus dual EnKF*, *hybrid EnKF versus dual EnKF*). The threshold is set to 25 %. For ensemble size 250, there is the significant case *local EnKF versus damped EnKF* for 15 % difference in mean RMSE, but there are also three comparisons with differences 16 %, 17 % and 19 % that are not significant (*iterative EnKF versus dual EnKF*, *hybrid EnKF versus dual EnKF*, and *local EnKF versus dual EnKF*). As for the tracer setup, the smallest significant RMSE difference is explained by the relatively narrow distribution of RMSEs for the damped EnKF (cf. Figure 4.7). The threshold is set to 18 %. For 10 synthetic experiments and ensemble size 50, there are two differences of 9 %, one is significant (*local EnKF versus hybrid EnKF*) and one is not (*classical EnKF versus dual EnKF*). Thus, the threshold is set to 9 %. For ensemble size 250, there is a difference of 5 % that is significant (*local EnKF versus iterative EnKF*) and there is one of 6 % that is not (*damped EnKF versus dual EnKF*). The threshold is set to 5 %. For 100 synthetic experiments and ensemble size 50, there is a difference of 4 % that is significant (*iterative EnKF versus classical EnKF*) and there is one of 3 % that is not (*hybrid EnKF versus normal score EnKF*). The threshold is set to 3 %. For ensemble size 250, there is a difference of 2 % that is significant (*hybrid EnKF versus iterative EnKF*) and there is another one of 2 % that is not (*normal score EnKF versus classical EnKF*). The threshold is set to 2 %. All thresholds are provided in Table 4.1.

4.3.3 Large ensembles

Results from synthetic experiments for ensemble sizes 500, 1 000 and 2 000 are shown in Figure 4.6. RMSE means are computed from 100 synthetic experiments instead of the 1 000 synthetic experiments that were calculated for the smaller ensemble sizes. This leads to a larger

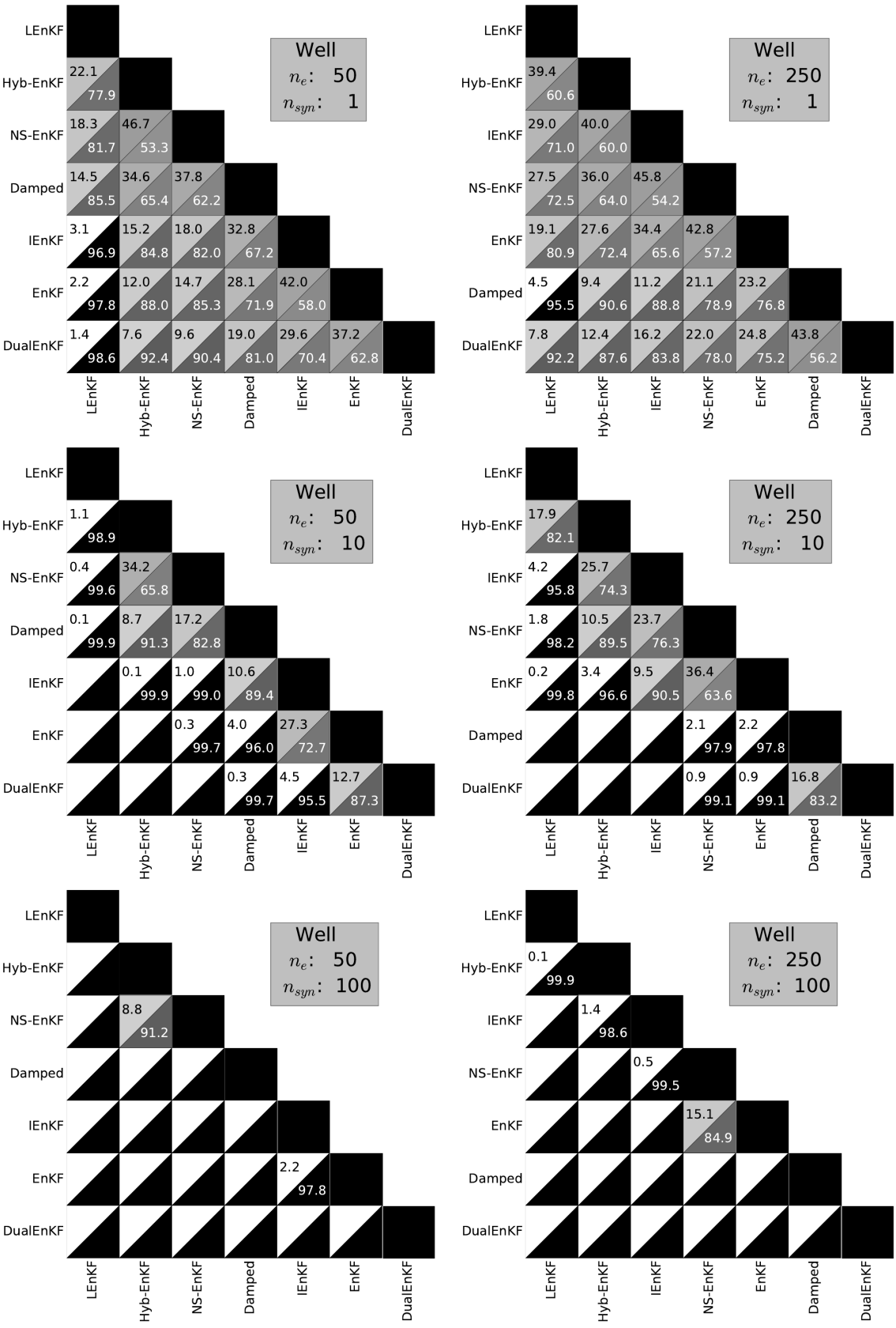

Figure 4.5: The same as Figure 4.3 for the well model. RMSE = root-mean-square error

Table 4.1: RMSE differences in % which are considered to be significant for the two simulation set-ups as a function of the number of synthetic experiments n_{syn} and the number of ensemble members n_e.

	n_{syn}	$n_e = 50$	$n_e = 70$	$n_e = 100$	$n_e = 250$	$n_e = 500$	$n_e = 1\,000$	$n_e = 2\,000$
Tracer	1	> 16	> 16	> 16	15	13	15	14
	10	9	9	8	6	6	5	4
	100	2	3	3	2	-	-	-
Well	1	25	20	23	18	17	16	15
	10	9	9	8	5	5	5	5
	100	3	< 3	2	2	-	-	-

uncertainty in the mean calculation. Additionally, for ensemble sizes 500, 1 000 and 2 000, RMSE means are very similar. RMSE means are even not always smaller for a larger ensemble size (e.g., example for Dual EnKF and the tracer model, the RMSE mean for ensemble size 500 is smaller than the RMSE mean for ensemble size 1 000). The ranking of the performance of the different EnKF variants for the large ensemble sizes is similar to the method ranking for ensemble size 250. In addition, for larger ensemble sizes the number of synthetic experiments that is needed to show significant difference in performance of EnKF methods is only marginally smaller than for ensemble size 250 (see Table 4.1).

4.3.4 RMSE distributions

The mean RMSE distributions calculated for the two model setups, $n_{syn} = 1, 10$, or 100 and ensemble sizes 50 and 250 are displayed in Figure 4.7. For $n_{syn} = 10$ or 100, the mean RMSE distributions are close to Gaussian, but for $n_{syn} = 1$ the spread is larger with more outliers. The narrow distributions calculated on the basis of 100 synthetic experiments, for the different EnKF methods, show little overlap. Comparing RMSE distributions for ensemble sizes 50 and 250, it can be seen that there are less outliers for the larger ensemble size. If I compare the RMSE distributions for the tracer and well model, it is clear that especially the damped EnKF has a different distribution from the rest. Whereas it has the narrowest distributions for the tracer setup, it has one of the widest distributions for the well setup.

The width of the RMSE distributions illustrates the variability of the estimation results related to the random seed variation. The large overlap of RMSE distributions for the different EnKF variants suggests that the estimation result is more dependent on the random seed than on the choice of EnKF method. If I base the analysis on 10 or 100 synthetic experiments, the distributions get narrower allowing to detect significant differences in performance of different EnKF methods.

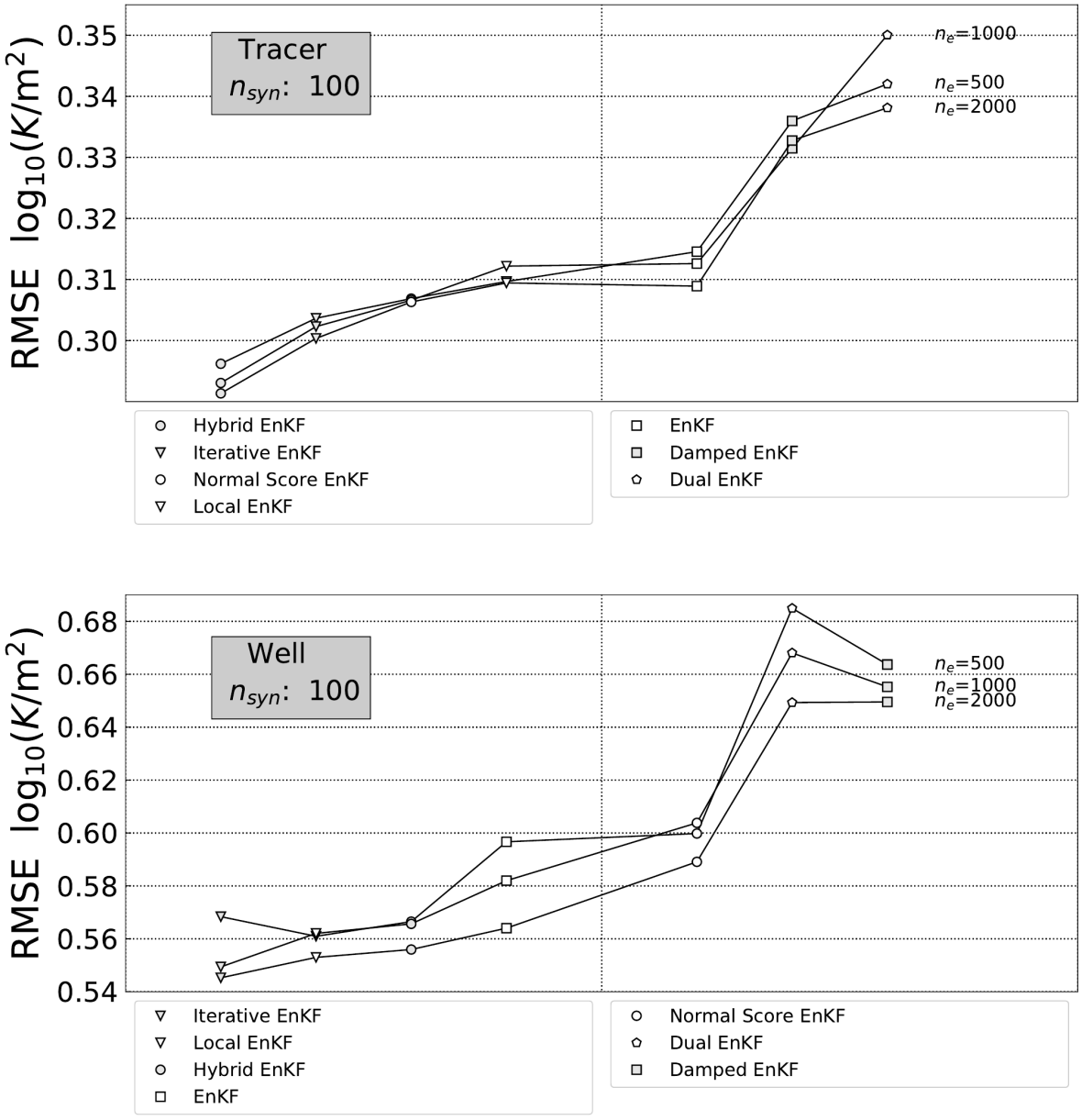

Figure 4.6: Mean RMSEs for seven EnKF methods derived from 100 synthetic experiments for the tracer and well model and ensemble sizes 500, 1 000 and 2 000.

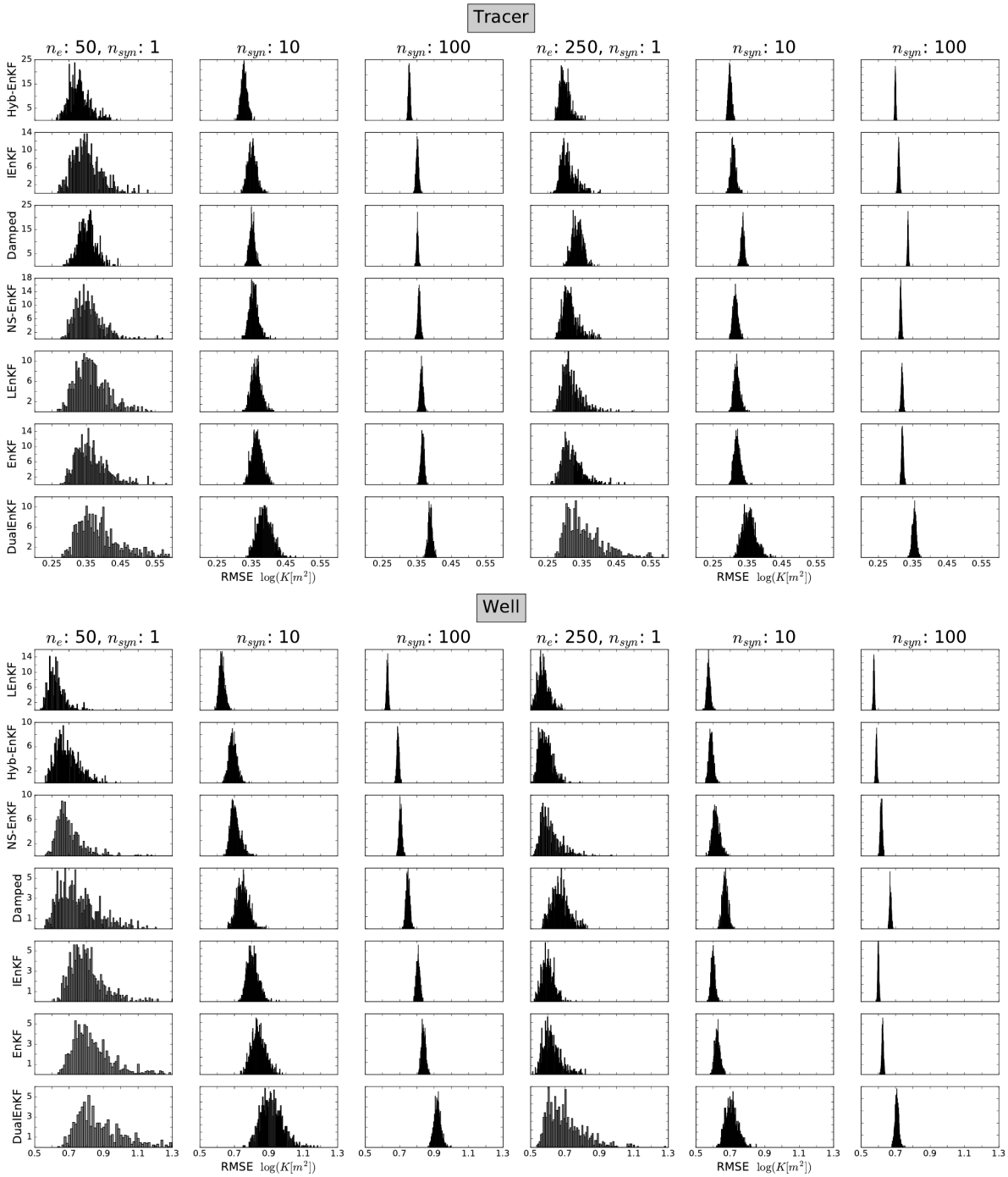

Figure 4.7: Mean root-mean-square error (RMSE) distributions for ensemble sizes 50 and 250, for 1, 10 and 100 synthetic experiments, for the tracer and well model. EnKF = ensemble Kalman filter.

4.3.5 Variation of EnKF method parameters

For the local EnKF and the hybrid EnKF, the correlation length λ and mixing parameter β were varied, respectively. The resulting mean RMSEs are displayed in Figure 4.8. For the tracer setup and ensemble size 50, the local EnKF with the smallest correlation length of 25 m yields the smallest mean RMSE. On the contrary, for larger ensemble sizes the smallest correlation length yields the largest mean RMSE and a correlation length of 150 m results in the smallest mean RMSE. For the well model, the correlation length of 150 m yields the smallest mean RMSEs for all ensemble sizes. Although the results are affected by the parameter values, the performance of the local EnKF is not so strongly influenced by the choice of the correlation length in these cases, except for ensemble size 250, for which a small correlation length results in a large RMSE affecting the ranking of the EnKF methods.

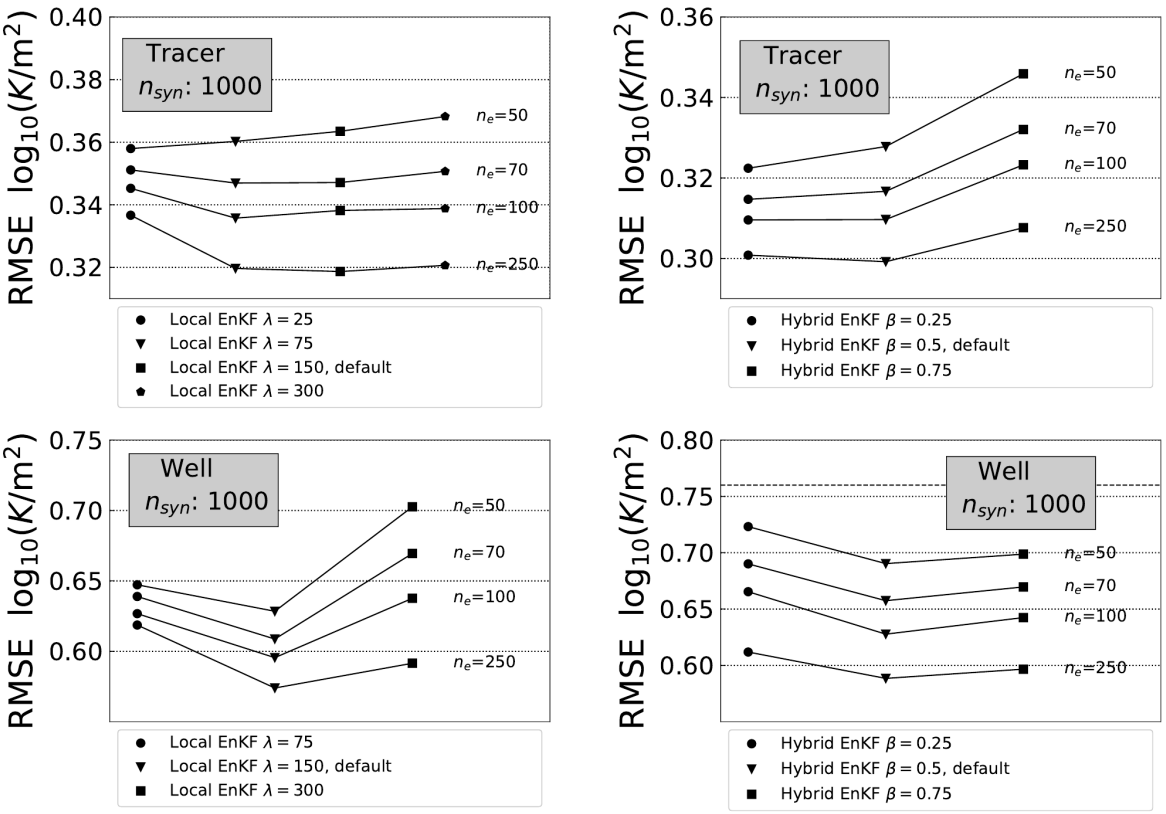

Figure 4.8: Variation of mean root-mean-square error (RMSE) for the local EnKF and the hybrid EnKF, as a function of values for the parameters λ and β and ensemble size. Results are shown for both the tracer model and the well model.

For the hybrid EnKF, the parameter β also has a certain impact on the performance, which is larger for small ensemble sizes. This influences the ranking of the methods, but to a limited extent. Even though the influences of specific parameter settings in these computations are limited, these examples do show the general importance of parameter settings which add an additional uncertainty component to model comparisons.

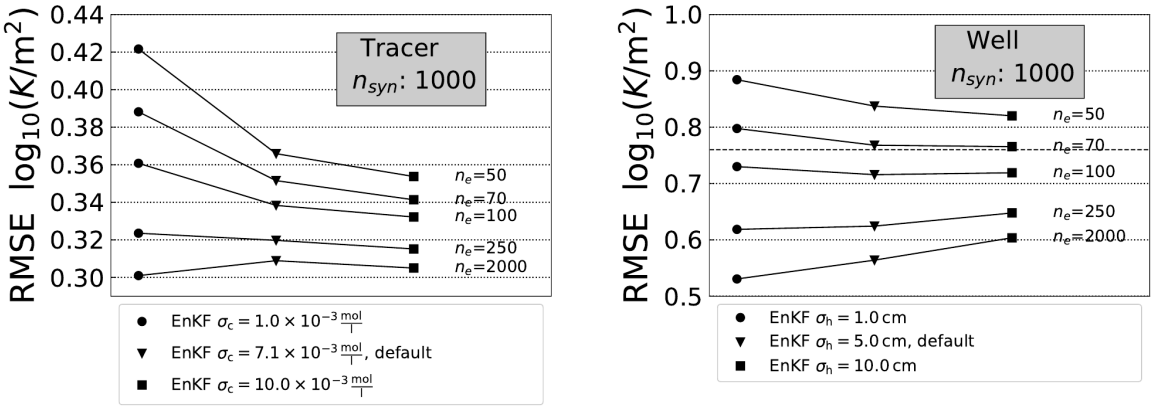

Figure 4.9: Variation of mean root-mean-square error (RMSE) for different observation noises in classical ensemble Kalman filter (EnKF), as function of ensemble size. Results are shown for both the tracer model and the well model.

4.3.6 Varying the observation noise

To illustrate the effect of observation noises, the tracer and well setups were computed for classical EnKF and observation noises larger and smaller than the noise level chosen for the default synthetic experiment. In a data assimilation using field observations, the observation noise would be defined by the type of measurement. However, in synthetic experiments from this work the observation noise is set somewhat arbitrarily. Thus, in this synthetic setting it makes sense to check the influence of varying this synthetic observation noise. In Figure 4.9, the resulting RMSE means are shown for ensemble sizes 50, 70, 100, 250 and 2 000. For ensemble size 50, in both setups, smaller observation noise leads to larger RMSE means, whereas the larger observation noise leads to smaller RMSE means. This trend changes, when the ensemble size is larger. For ensemble size 2 000, in both setups, the smallest observation noise results in the smallest RMSE mean.

Results for different observation noises can be understood taking into account two different impacts of observation noise on the simulation results. First, smaller observation noise results in a higher weight for observations, that is, a stronger correcting influence of the observations. It should result on average in smaller RMSEs as long as its influence is correctly weighted in the filter and it is therefore important that model covariances are correct, which is approximately the case if the ensemble size is large. For small ensemble sizes, model covariances differ more strongly from the true values resulting in (1) an incorrect weighting and (2) the risk that updates go in the wrong direction resulting in larger RMSEs.

The default values for the observation noise were originally chosen because they represent the behavior of typical measurement devices. According to these results, they also seem to constitute a sensible trade-off between filter divergence due to unjustifiably small observation noise and information loss due to unjustifiably large observation noise.

4.3.7 Discussion and Conclusion

The tracer model and the well model differ in a number of properties, including boundary conditions, correlation lengths of the reference permeability fields, measurement types, and the number of measurements that are assimilated. The data assimilation experiments for each of these two model setups result in different RMSE values. Nevertheless, for the two model setups similar conclusions are reached in terms of the number of synthetic experiments which is needed to show that one EnKF variant outperforms another one. Given these results, I recommend that for comparisons of EnKF variants 10 or more synthetic experiments are needed. A number of 10 synthetic experiments is needed to show that one EnKF variant significantly outperforms another EnKF variant if the two methods show a difference in RMSE of at least 10 %. It can be argued that a RMSE difference of 10 % is important enough to be detected. For really small RMSE differences of 2 % around 100 synthetic experiments are needed.

The study in this chapter only considers variations in the initial ensemble of rock permeabilities. However, in reality one faces other sources of uncertainty like additional uncertain parameters, model forcings, and boundary conditions, which affect the study outcome. Therefore, I feel that the main conclusion of this chapter, the need for 10 or more synthetic experiments to compare different EnKF variants, is not overly pessimistic, taking into account that several other factors also determine the relative performance of methods.

If I focus on the ranking of the EnKF methods, I find that the physical model setup and measurement type have a strong influence on the ranking of the EnKF methods. The synthetic experiments featuring two tracer measurements lead to permeability fields that were closer to the corresponding synthetic true than the synthetic experiments featuring 48 head measurements. The magnitude of the observation noise, which should, in principle, be determined by the measurements, was also observed to influence filter performance. The degree of heterogeneity of the permeability field might also influence the relative performance of a given EnKF variant. For example, the iterative EnKF might perform relatively better (compared to other EnKF variants) for strongly heterogeneous permeability fields as non-linearity is more prominent for those fields. On the other hand, for strongly non-Gaussian permeability fields it is expected that the normal score EnKF will improve its relative performance compared to other methods. For the local EnKF and the hybrid EnKF, an influence of the parameter choice on the filter performance was observed.

The ranking of the methods differed between the two physical model setups, emphasizing even more that the comparison of EnKF variants is not straightforward and not only affected by random fluctuations related to the synthetic case but also by other factors. It needs to be stressed that therefore no attempt is made to rank the EnKF variants, but to show the impact of random factors and also the physical model setup on the comparison of EnKF variants.

si parva licet componere magnis

Georgica (Verg. georg. IV, 175)
VERGIL

5

Pilot point ensemble Kalman filter

Spurious correlations are the main unwanted uncertainties in permeability estimation using the ensemble Kalman filter (EnKF). The pilot point ensemble Kalman filter (PP-EnKF) can remedy spurious correlations.

This chapter is organized as follows. After a small introduction, the Kalman filter and EnKF are introduced in a slightly different form compared to the introduction of this thesis, thereby preparing the subsequent introduction of the PP-EnKF. Then, the PP-EnKF is compared to the hybrid EnKF and the local EnKF, two methods manipulating the covariance matrix in the EnKF update in order to reduce spurious correlations. The comparison setup is explained, focusing on the differences that arise compared to Chapter 4 by introducing the PP-EnKF into the method comparison. The results of the performance comparison are presented, including three performance tests specifically designed for the PP-EnKF: (1) performance comparison for varied prior correlation length, (2) performance comparison for pilot point grid variation, and (3), most importantly, a comparison of spurious correlation reduction for the PP-EnKF compared to the EnKF. Finally, the conclusions of this chapter are presented.

Contents

adapted from | Keller, J., Hendricks Franssen, H., Nowak, W., Investigating the Pilot Point Ensemble Kalman Filter for geostatistical inversion and data assimilation, Advances in Water Resources, 155, 104010 (2021).

5.1 Introduction

For many EnKF methods, sampling errors of the cross-covariances between dynamic variables and parameters are responsible for filter divergence. A number of EnKF methods have been proposed and applied to tackle such shortcomings of the EnKF related to small ensemble sizes. Two examples are the local EnKF (Hamill et al., 2001) and the hybrid EnKF (Hamill and Snyder, 2000). In the local EnKF (also called covariance localization), covariances are restricted to the vicinity of observations. In the hybrid EnKF, the full covariance matrix is a weighted sum of the ensemble covariance matrix, which is the covariance of the classical Kalman filter update, and a fixed covariance matrix, which is defined before the assimilation. Additional EnKF methods include the damped EnKF (Hendricks Franssen and Kinzelbach, 2008), and the iterative EnKF (Sakov et al., 2012). In the damped EnKF, the proposed EnKF update is multiplied by a damping factor to reduce the update of the parameters. This approach reduces problems with filter inbreeding. The iterative EnKF restarts the whole assimilation after every EnKF update.

Further developments of methods related to the EnKF include the application of an ensemble smoother (Evensen and van Leeuwen, 2000, Chen and Oliver, 2011, Cosme et al., 2012, Emerick and Reynolds, 2013), and in particular an iterative version of this smoother (Bocquet and Sakov, 2013). Additionally, the iterative EnKF has been tested for additive model error (Sakov et al., 2018). A disadvantage of iterative approaches and smoothers is the strong increase in needed compute time. Also, adaptive covariance inflation of the EnKF has been extensively tested recently (Raanes et al., 2019). Another line of research focuses on the model equations and reducing computational effort by combining EnKF variants with methods such as principal

component analysis and reduced basis (Kang et al., 2017, Pagani et al., 2017, Xiao et al., 2018). However, these methods do not tend to reduce problems with spurious correlations and filter inbreeding.

In this chapter, I mathematically derive and extensively test the pilot point ensemble Kalman filter (PP-EnKF) as an alternative approach for parameter estimation. An EnKF method using pilot points that may be similar to the PP-EnKF is sketched and tested in a small suite of four tests by Heidari et al. (2013), using a 50 member ensemble in a petroleum reservoir model. The result was that this method yielded larger RMSEs than the classical EnKF, getting better as the number of pilot points approached the full set of parameters. On the other hand, spatial variability was found to be better preserved by the new method compared to the classical EnKF. In a second performance comparison, Tavakoli et al. (2013) tested the method from Heidari et al. (2013), called EnKF-PP by Tavakoli et al. (2013) and possibly quite similar to the method presented in this chapter, against the EnKF, ensemble smoothers and null-space Monte Carlo methods using a 200 member ensemble in a multi-phase model. In this comparison, the EnKF-PP performed well, yielding the smallest RMSE for the estimated logarithmic permeability field. In Tavakoli et al. (2013), the spread of the estimated EnKF-PP field was actually smaller than for the other ensemble methods, a result that is contradictory to the findings of Heidari et al. (2013). This makes it interesting to give a mathematically rigorous derivation and investigation of the performance of the PP-EnKF from this work. Interesting applications of pilot points in inverse modeling beyond two-point statistics can be found in Li et al. (2013) and Cao et al. (2018). In these works, pilot points are used with a pattern-based inverse method making the method faster and additionally improving the parameter estimation.

The specific contribution of the work in this chapter is twofold. First, the explicit mathematical derivation of the PP-EnKF, and secondly, an extensive testing of the PP-EnKF in a large comparison. The original publication (Heidari et al., 2013) contains no mathematical derivation, while the derivation in this work starts at the classical Kalman filter, shows the differences of the classical Kalman filter and a pilot point Kalman filter and, finally, translates the derivation to the corresponding ensemble methods obtaining the full PP-EnKF filter equations and their ideal (linearized) estimation variance. This derivation provides a rigorous mathematical and statistical foundation to the method, and provides insights on its behavior in various situations. Regarding the testing, the performance tests in this thesis of the PP-EnKF are based on a much larger test set compared to earlier tests in Heidari et al. (2013) and Tavakoli et al. (2013). Statistically independent repetition of performance tests includes RMSEs and overall standard deviations from 1 000 synthetic experiments and, additionally, the comparison of full correlation fields from 10 synthetic experiments. The latter comparison is used to assess the ability of the PP-EnKF to suppress spurious correlations.

In this study, the ensemble equations of the PP-EnKF are introduced by first defining the simpler equations of the pilot point Kalman filter (PP-KF). The PP-KF is obtained by splitting the state vector of the Kalman filter into dynamic variables, pilot point parameters and non-pilot point parameters. The Kalman update is applied to dynamic variables and pilot point parameters. The remaining parameter values are updated by a kriging interpolation of the update. In the

next step, the ensemble versions of the equations are formulated, resulting in the PP-EnKF. I show how the update of the PP-EnKF, which is a mixture of an update and an interpolation of this update, differs from the update in the classical EnKF. Additionally, I compare the PP-EnKF to the local EnKF and the hybrid EnKF to show that, while aiming at the same goal of suppressing spurious correlations, the PP-EnKF provides a useful alternative.

The performance of the PP-EnKF is evaluated by comparing it to seven other EnKF methods (damped, iterative, local, hybrid, dual (Moradkhani et al., 2005), normal score (NS-EnKF, Journel and Huijbregts, 1978, Goovaerts, 1997, Zhou et al., 2011, Schöniger et al., 2012, Li et al., 2012) and the classical EnKF). The comparison procedure is adopted from Chapter 4. The performance evaluations are carried out for two parameter estimation setups: a 2D tracer transport problem and a 2D flow problem with one injection well. The estimated parameter field in these setups is a permeability field. In each setup, 1 000 synthetic experiments are computed for the PP-EnKF and for the seven other EnKF methods, and each one for ensemble sizes of 50, 70, 100, and 250. In each synthetic experiment, I assess both the updated parameter fields and their ensemble variance. The ensemble variance is compared to a reference EnKF estimation with a large ensemble size of 10 000. Next to the RMSE comparison, there is a comparison of the correlation fields driving the Kalman update using 10 synthetic experiments for ensemble size 50. The correlation output of these synthetic experiments is compared to the correlation field of the reference EnKF with ensemble size 10 000 by computing the RMSE difference of the correlation fields. Additionally, the impact of the correlation length of the prior permeability fields is investigated. The correlation length is important in the PP-EnKF, because it not only affects the prior permeability field but also the kriging interpolation. Finally, to check the influence of the pilot point grid on the performance of the PP-EnKF, different grid configurations are tested.

5.2 Pilot point ensemble Kalman filter

Before the PP-EnKF is presented, I quickly recall the equations for the ensemble Kalman filter (Kalman, 1960, Evensen, 1994), because the PP-EnKF differs from the EnKF only in the update equation. A full account of the EnKF can be found in Evensen (2003).

5.2.1 Ensemble Kalman filter

For the equations of the classical Kalman filter, please consult Section 2.2.1. In particular, the following four equations will be reformulated in the next section: (1) The forward equations of the Kalman filter, Equation (2.12) for the state vector $\mathbf{x}^f$ and Equation (2.13) for the covariance matrix $\mathbf{P}$; and (2) the update equations of the Kalman filter, Equation (2.14) for the state vector $\mathbf{x}^a$ and Equation (2.15) for the covariance matrix $\mathbf{P}^a$.

I now turn to the EnKF. The n_e forward equations of the EnKF are given by

$$\mathbf{x}_i^f = M\left(\mathbf{x}_i^0\right), \qquad i \in \{1, \cdots, n_e\}, \tag{5.1}$$

where n_e is the size of the ensemble of state vector realizations. The state vector realizations $\mathbf{x}_i^0 \in \mathbb{R}^{n_s}$ make up the initial state vector ensemble. The state vector realizations $\mathbf{x}_i^f$ make up the predicted state vector ensemble. The forward operator M can be non-linear, such as a numerical solution of the groundwater flow equation. All covariance matrices in the EnKF are computed from the ensemble of state vector realizations. The predictions $\mathbf{x}_i^f$ serve as input to the update equation of the EnKF.

The EnKF update equation (2.20) is given by

$$\mathbf{x}_i^a - \mathbf{x}_i^f = \mathbf{P}_e\mathbf{H}^T \left(\mathbf{H}\mathbf{P}_e\mathbf{H}^T + \mathbf{R}\right)^{-1} \left(\mathbf{d}_i - \mathbf{H}\mathbf{x}_i^f\right), \qquad i \in \{1, \cdots, n_e\}. \tag{5.2}$$

This equation is very similar to the Kalman filter update of the mean state vector in Equation (2.14), with two important differences. First, Equation (5.2) is computed n_e times for the realizations $\mathbf{x}_i^f$ and $\mathbf{x}_i^a$ of the state vector, and realizations $\mathbf{d}_i$ of the measurements. Secondly, the covariance matrix $\mathbf{P}_e$ is estimated from the ensemble of state vector realizations. Therefore, an analogue of update equation (2.15) for the covariance matrix is not needed for the EnKF.

In this chapter, all Kalman update equations are written in terms of the difference between an estimated quantity and the same quantity computed from the forward simulation. When the Kalman filter update or the EnKF update is mentioned, this refers to either this difference or the formula for its computation. Expressing the update equation in this form is convenient for the introduction of the pilot point EnKF (PP-EnKF). In the PP-EnKF, the updates on the left-hand sides of Equations (2.14), (2.15), and (5.2) are the input of a kriging interpolation (Deutsch and Journel, 1995).

5.2.2 Pilot point EnKF

Retracing the update equations in the last section, I will now introduce the update equations of the pilot point Kalman filter (PP-KF) and, subsequently, the update equations of the pilot point ensemble Kalman filter (PP-EnKF). Here, the explicit set of equations of the PP-EnKF is rigorously derived and thoroughly discussed.

The main idea of the PP-EnKF is to update in a first step only parameters at a fixed subset of locations, called pilot points (RamaRao et al., 1995, Gómez-Hernández et al., 1997). The positions of the pilot points are defined before the PP-EnKF starts. In the rest of the model domain apart from the pilot points, parameter values are initialized by random geostatistical simulation and are then subject to an interpolated update. The interpolation of the update is generated by ordinary kriging of the pilot point updates and added to the geostatistically simulated fields. I use ordinary kriging, since I do not assume a background trend as in universal kriging, or a correlation with other variables as in cokriging. The update of a certain parameter is defined as the difference between the parameter value after the EnKF update step and before the EnKF update step. While the updates apart from pilot points are interpolations of the perturbations calculated at the pilot points, the full parameter fields retain their initial spatial variability.

In general, two covariance structures are important for the algorithm of the PP-EnKF. First, there is the covariance structure used by the kriging interpolation. This covariance structure is defined a priori and is reused for the interpolation at each update step. Secondly, there is the covariance structure used for initial geostatistical simulation. In the PP-EnKF, both covariance structures are equal and defined according to prior knowledge. Overall, the partially fixed covariance structures should enable the PP-EnKF to suppress spurious correlations, while still providing ensemble-based updates that potentially affect the whole model domain. I argue that these features make the PP-EnKF a useful alternative to existing EnKF methods such as the hybrid EnKF and the local EnKF.

To define the update of the PP-KF, the state vector and its covariance matrix are split into three parts, the pilot point parameters, the non-pilot point parameters, and the dynamic variables:

$$\mathbf{x}^f = \begin{pmatrix} \mathbf{x}_p^f \\ \mathbf{x}_r^f \\ \mathbf{x}_d^f \end{pmatrix} \in \mathbb{R}^{n_s} \qquad \mathbf{P} = \begin{pmatrix} \mathbf{P}_{pp} & \mathbf{P}_{pr} & \mathbf{P}_{pd} \\ \mathbf{P}_{rp} & \mathbf{P}_{rr} & \mathbf{P}_{rd} \\ \mathbf{P}_{dp} & \mathbf{P}_{dr} & \mathbf{P}_{dd} \end{pmatrix} \in \mathbb{R}^{n_s \times n_s} \ . \tag{5.3}$$

Here, $\mathbf{x}_p^f \in \mathbb{R}^{n_p}$ contains parameter values at pilot point positions, and $n_p \in \mathbb{N}$ is the number of pilot point parameters. $\mathbf{x}_r^f \in \mathbb{R}^{n_r}$ denotes the parameters at non-pilot point locations, where $n_r \in \mathbb{N}$ is the number of non-pilot point parameters. Finally, $\mathbf{x}_d^f \in \mathbb{R}^{n_d}$ contains the $n_d \in \mathbb{N}$ dynamic variables. The sum $n_s = n_p + n_r + n_d$ is the total number of entries in the state vector. The covariance matrix $\mathbf{P}_{pp} \in \mathbb{R}^{n_p \times n_p}$ contains the covariances between all pairs of parameters at pilot points. $\mathbf{P}_{pr} \in \mathbb{R}^{n_p \times n_r}$ and $\mathbf{P}_{rp} \in \mathbb{R}^{n_r \times n_p}$ contain covariances between parameters at pilot points and parameters at non-pilot points. $\mathbf{P}_{rr} \in \mathbb{R}^{n_r \times n_r}$ contains covariances between pairs of parameters at non-pilot points. $\mathbf{P}_{dp} \in \mathbb{R}^{n_d \times n_p}$ and $\mathbf{P}_{pd} \in \mathbb{R}^{n_p \times n_d}$ contain covariances between dynamic variables and parameters at pilot points. $\mathbf{P}_{dr} \in \mathbb{R}^{n_d \times n_r}$ and $\mathbf{P}_{rd} \in \mathbb{R}^{n_r \times n_d}$ contain covariances between dynamic variables and parameters at non-pilot points. Finally, $\mathbf{P}_{dd} \in \mathbb{R}^{n_d \times n_d}$ contains covariances between pairs of dynamic variables.

The update of the PP-KF consists of two steps: updating the restricted state vector and then interpolating the update. First, the pilot point update equations are discussed. The form of the mean vector update equation is identical to Equation (2.14) of the Kalman filter, restricted to dynamic variables and pilot-point parameters:

$$\begin{pmatrix} \mathbf{x}_p^a \\ \mathbf{x}_d^a \end{pmatrix} - \begin{pmatrix} \mathbf{x}_p^f \\ \mathbf{x}_d^f \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{pp} & \mathbf{P}_{pd} \\ \mathbf{P}_{dp} & \mathbf{P}_{dd} \end{pmatrix} \mathbf{H}^T \left(\mathbf{P}_{yppy} + \mathbf{R} \right)^{-1} \left(\mathbf{d} - \mathbf{H} \begin{pmatrix} \mathbf{x}_p^f \\ \mathbf{x}_d^f \end{pmatrix} \right) \ . \tag{5.4}$$

Definitions and explanations for the PP-KF versions of the covariance matrix among simulated measurements $\mathbf{P}_{yppy} \in \mathbb{R}^{n_m \times n_m}$ and the measurement operator $\mathbf{H} \in \mathbb{R}^{n_m \times (n_p + n_d)}$ can be found in Section 5.2.3.

The PP-KF update of the covariance matrix has the same form as Equation (2.15), but again is

restricted to dynamic variables and parameters at pilot points:

$$\begin{pmatrix} \mathbf{P}^a_{pp} & \mathbf{P}^a_{pd} \\ \mathbf{P}^a_{dp} & \mathbf{P}^a_{dd} \end{pmatrix} - \begin{pmatrix} \mathbf{P}_{pp} & \mathbf{P}_{pd} \\ \mathbf{P}_{dp} & \mathbf{P}_{dd} \end{pmatrix} = - \begin{pmatrix} \mathbf{P}_{pp} & \mathbf{P}_{pd} \\ \mathbf{P}_{dp} & \mathbf{P}_{dd} \end{pmatrix} \mathbf{H}^T \left(\mathbf{P}_{yppy} + \mathbf{R} \right)^{-1} \mathbf{H} \begin{pmatrix} \mathbf{P}_{pp} & \mathbf{P}_{pd} \\ \mathbf{P}_{dp} & \mathbf{P}_{dd} \end{pmatrix} . \tag{5.5}$$

$\mathbf{x}^a_p \in \mathbb{R}^{n_p}$ and $\mathbf{P}^a_{pp} \in \mathbb{R}^{n_p \times n_p}$ are the updates of mean vector and covariance matrix of the parameters at pilot point locations, $\mathbf{x}^a_d \in \mathbb{R}^{n_d}$ and $\mathbf{P}^a_{dd} \in \mathbb{R}^{n_d \times n_d}$ are the corresponding quantities for the dynamic variables. $\mathbf{P}^a_{dp} \in \mathbb{R}^{n_d \times n_p}$ and $\mathbf{P}^a_{pd} \in \mathbb{R}^{n_p \times n_d}$ contain updated covariances between pilot-point parameters and dynamic variables.

Now I discuss the second step, i.e., the transfer of the update to non-pilot point parameters by kriging interpolation. Through the kriging interpolation, the updates from Equations (5.4) and (5.5) are used to compute updates for means and covariances of parameters at non-pilot points:

$$\mathbf{x}^a - \mathbf{x}^f = P_p \left[\begin{pmatrix} \mathbf{x}^a_p \\ \mathbf{x}^a_d \end{pmatrix} - \begin{pmatrix} \mathbf{x}^f_p \\ \mathbf{x}^f_d \end{pmatrix} \right] \tag{5.6}$$

The interpolation operator P_p is defined as

$$P_p = \begin{pmatrix} 1 & 0 \\ \mathbf{P}^0_{rp} (\mathbf{P}_{pp})^{-1} & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{n_s \times (n_p + n_d)} . \tag{5.7}$$

Here, $\mathbf{P}^0_{rp} \in \mathbb{R}^{n_r \times n_p}$ is a fixed covariance matrix between pilot point parameters and non-pilot point parameters. This matrix is specified in advance. A suitable choice for $\mathbf{P}^0_{rp}$ are prior covariances equal to the covariances used in the generation of the prior permeability fields as these covariances reflect the prior knowledge. The interpolation equation for the covariance matrix is

$$\mathbf{P}^a - \mathbf{P} = P_p \left[\begin{pmatrix} \mathbf{P}^a_{pp} & \mathbf{P}^a_{pd} \\ \mathbf{P}^a_{dp} & \mathbf{P}^a_{dd} \end{pmatrix} - \begin{pmatrix} \mathbf{P}_{pp} & \mathbf{P}_{pd} \\ \mathbf{P}_{dp} & \mathbf{P}_{dd} \end{pmatrix} \right] (P_p)^T . \tag{5.8}$$

A closer discussion of the interpolation operator P_p can be found in Section 5.2.3.

The equations for the PP-EnKF are very similar to Equations (5.3), (5.4), and (5.6) for the mean update of the PP-KF. The state vector is split for every realization in the ensemble:

$$\mathbf{x}^f_i = \begin{pmatrix} \mathbf{x}^f_{p,i} \\ \mathbf{x}^f_{r,i} \\ \mathbf{x}^f_{d,i} \end{pmatrix} \in \mathbb{R}^{n_s} , \qquad i \in \{1, \cdots, n_e\} . \tag{5.9}$$

Then, the update of the state vector is calculated for each realization of pilot point parameters and dynamic variables:

$$\begin{pmatrix} \mathbf{x}^a_{p,i} \\ \mathbf{x}^a_{d,i} \end{pmatrix} - \begin{pmatrix} \mathbf{x}^f_{p,i} \\ \mathbf{x}^f_{d,i} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{pp,e} & \mathbf{P}_{pd,e} \\ \mathbf{P}_{dp,e} & \mathbf{P}_{dd,e} \end{pmatrix} \mathbf{H}^T \left(\mathbf{P}_{yppy,e} + \mathbf{R} \right)^{-1} \left(\mathbf{d}_i - \mathbf{H} \begin{pmatrix} \mathbf{x}^f_{p,i} \\ \mathbf{x}^f_{d,i} \end{pmatrix} \right) , \quad i \in \{1, \cdots, n_e\} . \tag{5.10}$$

Finally, the interpolation is applied to each realization:

$$\mathbf{x}_i^a - \mathbf{x}_i^f = P_p \left(\begin{pmatrix} \mathbf{x}_{p,i}^a \\ \mathbf{x}_{d,i}^a \end{pmatrix} - \begin{pmatrix} \mathbf{x}_{p,i}^f \\ \mathbf{x}_{d,i}^f \end{pmatrix} \right) , \qquad i \in \{1, \cdots, n_e\} . \tag{5.11}$$

There are two main differences between the updates of the PP-EnKF and the PP-KF. First, the update of the PP-EnKF consists of n_e equations, one for each realization of the ensemble. Secondly, all covariance matrices except $\mathbf{P}_{rp}^0$ are calculated from the ensemble of state vector realizations. When moving from the PP-KF to the PP-EnKF, initialization of the permeability field is randomized. Opposed to this randomization, the variability outside the pilot points, represented by the fixed matrix $\mathbf{P}_{rp}^0$, is chosen once and then fixed throughout the computation. In the next section, important matrices of the PP-KF and the PP-EnKF algorithms are explained in more detail.

5.2.3 Covariance matrix of the simulated observations and interpolation operator

The matrix $\mathbf{P}_{yppy}$ used in the update equations (5.4) and (5.5) of the PP-KF is an approximation of the covariance matrix of simulated measurement variables $\mathbf{P}_{yy} \in \mathbb{R}^{n_m \times n_m}$ appearing in the Kalman update (Equation (2.14)):

$$\mathbf{P}_{yy} = \mathbf{H}\mathbf{P}\mathbf{H}^T \tag{5.12}$$

$\mathbf{P}_{yppy}$ should approximate $\mathbf{P}_{yy}$ by taking into account only the dynamic variables and the pilot point parameters. The straightforward approximation with this property is obtained by removing the covariances that contain information from the non-pilot point locations, i.e. $\mathbf{P}_{rp}$, $\mathbf{P}_{pr}$, $\mathbf{P}_{rr}$, $\mathbf{P}_{dr}$, and $\mathbf{P}_{rd}$, from $\mathbf{P}$ before applying $\mathbf{H}$.

$$\mathbf{P}_{yppy} = \mathbf{H} \begin{pmatrix} \mathbf{P}_{pp} & 0 & \mathbf{P}_{pd} \\ 0 & 0 & 0 \\ \mathbf{P}_{dp} & 0 & \mathbf{P}_{dd} \end{pmatrix} \mathbf{H}^T \tag{5.13}$$

Note that $\mathbf{P}_{yppy} = \mathbf{P}_{yy}$ if $\mathbf{H}$ is a linear operator depending only on pilot point parameters and dynamic variables. A simple example of such a measurement operator $\mathbf{H}$ arises when all measurement variables are either dynamic variables or parameters at pilot points. For the PP-EnKF, the above remains true, with the amendment that the covariance matrices $\mathbf{P}_{yppy,e}$, $\mathbf{P}_{yy,e}$, $\mathbf{P}_{pp,e}$, $\mathbf{P}_{pd,e}$, $\mathbf{P}_{dp,e}$, and $\mathbf{P}_{dd,e}$ are calculated from an ensemble of realizations.

Now the interpolation operator P_p is discussed. The operator P_p defines the kriging interpolation, the second part of the update of both the PP-KF and the PP-EnKF. It determines the updates of non-pilot point parameters based on the updates of pilot point parameters. The definition of P_p was given in Equation (5.7), and it makes use of two covariance matrices, $\mathbf{P}_{pp}$ and $\mathbf{P}_{rp}^0$. The covariance matrix of the pilot point parameters $\mathbf{P}_{pp}$ is obtained from the forward simulation. Thus, the knowledge of $\mathbf{P}_{pp}$ originates from a mixture of the initially set prior covariance matrix and of the filter equations that were applied before the current update. In contrast, the covariance between the non-pilot point parameters and the pilot point parameters $\mathbf{P}_{rp}^0$ is fixed throughout

the computation. The superscript zero emphasizes two characteristic features of $\mathbf{P}^0_{rp}$, first, $\mathbf{P}^0_{rp}$ is fixed and, secondly, its elements may be prior covariances. Fixing $\mathbf{P}^0_{rp}$ is an integral part of the PP-EnKF, because this fixed covariance is one of the two reasons why I expect the PP-EnKF to suppress spurious correlations between pilot point parameters and non-pilot point parameters, the other reason being the reduction of the number of parameters in the Kalman update.

5.2.4 Comparison to the classical EnKF and other EnKF methods

I compare the update equations (5.10) and (5.11) of the PP-EnKF to the update equation (5.2) of the classical EnKF. If the measurement operator $\mathbf{H}$ operates on parameters at non-pilot points, this particular measurement information will be lost through the approximation in the PP-EnKF update, as shown in Section 5.2.3. This is a hypothetical case, because one can always locate pilot points at measurement locations. If the measurement operator $\mathbf{H}$ operates only on parameters at pilot points and dynamic variables, the main difference between the update equations is the kriging interpolation of the PP-EnKF. $\mathbf{P}_e$ in the update equation (Equation (5.2)) of the EnKF can be partitioned the same way as in Equation (5.3):

$$\mathbf{P}_e = \begin{pmatrix} \mathbf{P}_{pp,e} & \mathbf{P}_{pr,e} & \mathbf{P}_{pd,e} \\ \mathbf{P}_{rp,e} & \mathbf{P}_{rr,e} & \mathbf{P}_{rd,e} \\ \mathbf{P}_{dp,e} & \mathbf{P}_{dr,e} & \mathbf{P}_{dd,e} \end{pmatrix} \in \mathbb{R}^{n_s \times n_s}. \tag{5.14}$$

The middle column of this matrix does not influence the update equation, if the aforementioned restrictions on $\mathbf{H}$ hold. In the PP-EnKF, $\mathbf{P}_e$ is approximated by

$$P_p \begin{pmatrix} \mathbf{P}_{pp,e} & \mathbf{P}_{pd,e} \\ \mathbf{P}_{dp,e} & \mathbf{P}_{dd,e} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \mathbf{P}^0_{rp}(\mathbf{P}_{pp,e})^{-1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{P}_{pp,e} & \mathbf{P}_{pd,e} \\ \mathbf{P}_{dp,e} & \mathbf{P}_{dd,e} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{pp,e} & \mathbf{P}_{pd,e} \\ \mathbf{P}^0_{rp} & \mathbf{P}^0_{rp}(\mathbf{P}_{pp,e})^{-1}\mathbf{P}_{pd,e} \\ \mathbf{P}_{dp,e} & \mathbf{P}_{dd,e} \end{pmatrix}. \tag{5.15}$$

It is of interest to compare the two columns of matrix (5.15) to the first and last column of $\mathbf{P}_e$. Differences between the two matrices appear in the second row. The left columns of these matrices determine the update coming from measurements of parameters at pilot points. The updates of the EnKF and the PP-EnKF would be equal if the fixed covariance matrix $\mathbf{P}^0_{rp}$ would be equal to the ensemble covariance $\mathbf{P}_{rp,e}$ from the EnKF. Of course, the matrices will never be exactly equal due to sampling fluctuations. One of the positive effects of $\mathbf{P}^0_{rp}$ is that it does not suffer from spurious correlations.

The right column of matrix (5.15) determines the update coming from measurements of dynamic variables. Values at non-pilot points are updated according to $\mathbf{P}^0_{rp}(\mathbf{P}_{pp,e})^{-1}\mathbf{P}_{pd,e}$ instead of $\mathbf{P}_{rd,e}$. There is no direct covariance matrix between non-pilot points and dynamic variables in the PP-EnKF. Instead, the update is correlated with the pilot points and then interpolated according to $\mathbf{P}^0_{rp}$. Again, the main enhancement is the reduction of spurious correlations.

I now compare the update equations of the PP-EnKF to the update equations of two popular EnKF methods, the local EnKF and the hybrid EnKF. Similarly to the local EnKF and the

hybrid EnKF, the PP-EnKF uses a partially fixed covariance structure in order to suppress spurious correlations. In the local EnKF, updates are calculated directly for all parameters, but correlations between parameters (or correlations between dynamic variables and parameters) are set to zero if their distance to measurement locations exceeds a certain threshold. Like the local EnKF, the PP-EnKF aims to suppress spurious correlations. To achieve this, only the correlations between measurement variables and pilot points are calculated from the ensemble. I argue that the PP-EnKF could be preferred to the local EnKF in cases, when distant locations are significantly correlated. Correlations between these locations would be suppressed by the local EnKF, while the PP-EnKF includes them through correlations between distant pilot point locations.

In the hybrid EnKF, a mixture of covariance matrices is used in the update step, partially fixed and partially calculated from the ensemble. While this can yield good results, I argue that the PP-EnKF is an appealing alternative, because it interferes less with the statistics of the update step. The update step of the PP-EnKF is restricted to pilot point parameters and dynamic variables, but for these two sets of state vector variables the update is calculated exclusively from the ensemble-based correlations that are derived from the ensemble of forward model runs. The rest of the parameters is then updated by kriging of the updates at the pilot points, using the pre-defined fixed covariance matrix.

Implementing the PP-EnKF is relatively straightforward. In contrast to the hybrid or the local EnKF, there is no need to change the covariance matrix of the classical EnKF update equation. All changes of the PP-EnKF (compared to the classical EnKF) can be implemented by pre-processing (restricting the state vector to dynamic variables and parameters at pilot points) and post-processing (calculating the interpolated parameter values at non-pilot points). Thus, a modular implementation around an existing EnKF implementation is possible.

5.3 Design of the synthetic experiments

5.3.1 Subsurface models

The performance comparison of the PP-EnKF with other EnKF variants in this study is carried out for two transient subsurface model setups, first, a 2D solute transport model (*tracer model*), and, secondly, a 2D groundwater flow problem around an injection well (*well model*). The grids of both models consist of $31 \times 31 = 961$ identical squared cells. The size of the model domain for the tracer model is $62\,\mathrm{m} \times 62\,\mathrm{m}$; the size of the model domain for the well model is $620\,\mathrm{m} \times 620\,\mathrm{m}$. The simulation period is 1 200 days for the tracer model and 18 days for the well model. For both models, the simulation period is divided into 1 200 time steps. All forward simulations are computed with the numerical software SHEMAT-Suite (Rath et al., 2006, Clauser, 2003).

In the tracer model, a constant head difference of $11\,\mathrm{m} - 10\,\mathrm{m} = 1\,\mathrm{m}$ is implemented between the southern boundary and the northern boundary, accompanied by a constant concentration difference of $80 \times 10^{-3}\,\mathrm{mol/L} - 60 \times 10^{-3}\,\mathrm{mol/L} = 20 \times 10^{-3}\,\mathrm{mol/L}$. The remaining two boundaries

are impermeable. The initial conditions are a head of 10 m and a concentration of 60×10^{-3} mol/L throughout the model domain for the tracer model. In the well model, a head difference of 1 m is implemented between a central injection well location at coordinates (310 m, 310 m) and the four boundaries. Initial head is chosen as 10 m throughout the model domain for the well model. For both setups, standard properties of water as implemented by SHEMAT-Suite are used and the porosity of the background matrix is 10 %. The tracer is subjected to advective transport only.

Measurement locations vary between the two setups. The tracer model includes two measurement locations at coordinates (19 m, 31 m) and (43 m, 31 m). At these locations, tracer concentration and hydraulic head measurements are available at 100 evenly distributed times throughout the simulation period. The well model includes 49 measurement locations on a 7×7 grid throughout the model domain. At these locations, hydraulic head measurements are available at 60 evenly distributed times throughout the simulation period.

The synthetic reference permeability distributions for the two subsurface model setups are displayed in Figure 4.1 from the previous chapter, where the synthetic reference permeability distributions are introduced. The permeability fields are generated by sequential multi-Gaussian simulation (SGSIM, Deutsch and Journel, 1995). The same holds for the ensemble of prior permeability field realizations that is used to initialize the EnKF. A permeability mean of $-12.0 = \log_{10}(K/\mathrm{m}^2)$ is used for the synthetic reference and a mean of $-12.5 = \log_{10}(K/\mathrm{m}^2)$ is used for the prior permeability distributions. In both cases, the standard deviation is 0.5. The isotropic correlation length of the permeability fields is 50 m for the tracer model and 60 m for the well model.

5.3.2 Parameter settings for the PP-EnKF and other EnKF variants

Measurement noises are equal for all EnKF methods including the PP-EnKF. Measurement noises for the concentration measurements of the tracer model are set to $\sigma_{\mathtt{c}} = 7.1 \times 10^{-3}$mol/L and measurement noises for the hydraulic head measurements of both setups are set to $\sigma_{\mathtt{h}} = 5 \times 10^{-2}$ m. EnKF methods which require parameter settings include the damped EnKF, where the damping constant is set to 0.1 (Hendricks Franssen and Kinzelbach, 2008). In the local EnKF, the length scale (which is half the cutoff radius) is set to 150 m, which is larger than the correlation lengths. In the hybrid EnKF, the mixing constant is set to 0.5 and a diagonal background covariance matrix is specified. The damping factor of the damped EnKF and the parameter choices of the local EnKF and the hybrid EnKF were shown to yield the smallest root-mean-square errors in most synthetic experiments of the two setups of the previous performance comparison presented in Chapter 4.

The PP-EnKF needs as inputs the locations for the pilot points, and a covariance matrix for the kriging interpolation. I use 51 pilot points including all cell indices of the measurement locations of both setups (tracer model and well model). Thus, the pilot points lie on a 7×7 grid for both model setups with two additional pilot points in the center of the model corresponding to

the measurement locations of the tracer model. A regular grid of pilot points is documented to be beneficial (Capilla et al., 1997). The interpolation covariance $\mathbf{P}_{rp}^0$ is chosen identical to the prior covariance. This is implemented by computing the covariances in $\mathbf{P}_{rp}^0$ from 10 000 permeability fields, which are generated by SGSim with the same correlation length, mean and standard deviation as the prior permeability fields.

5.3.3 Performance comparison setup

In Chapter 4, seven EnKF variants have been compared (damped, iterative, local, hybrid, dual, normal score and classical). Now, I compare the PP-EnKF to these same seven methods by computing synthetic experiments for the two physical model setups introduced in the previous sections (compare Section 5.4.1). Additionally, correlation lengths of the initial permeability fields of both setups are varied to half and twice the correlation length of the synthetic truth (Section 5.4.3). This is done, because the correlation length plays a very prominent role in the PP-EnKF, more prominent than in the other EnKF methods, since it enters in the PP-EnKF not only as correlation length of the prior permeability field, but also as correlation length of the kriging interpolation. Thus, it is especially important to see how the PP-EnKF performs compared to other models, when it is subject to misspecified correlation lengths. One important degree of freedom in the PP-EnKF is the choice of the grid of pilot points. Thus, I compare the results of different sensible grid choices for the two model setups (Section 5.4.5). Next to the performance of the mean update, I am interested in the spatial variability of the permeability ensemble update by the PP-EnKF. This will help assess the ability of the PP-EnKF to suppress spurious correlations. To judge spatial variability, I compare the overall standard deviation of the ensembles of the updates of the various EnKF methods from all 1 000 synthetic experiments. Additionally, the full correlation fields of the EnKF and the PP-EnKF from 10 synthetic experiments are compared to a reference correlation field.

5.3.4 Performance comparison measures

Multiple synthetic experiments are needed to compare EnKF methods (see Chapter 4). Here I use root-mean-square errors (RMSEs) from 1 000 synthetic experiments to compare the PP-EnKF to the other seven EnKF methods. For each of the EnKF methods in the synthetic experiment j, the RMSE is computed as in Equation (4.15). Such a single RMSE is computed from the squared differences of estimated mean logarithmic permeabilities $\bar{\mathbf{Y}}_j \in \mathbb{R}^{n_g}$ and the synthetic reference $\mathbf{Y}_j^t \in \mathbb{R}^{n_g}$ across the n_g grid cells.

The RMSE measures the distance between the average over the estimated permeability field realizations and the synthetic true permeability field. Additionally, the overall standard deviation among realizations is introduced as a measure for the uncertainty of the estimated permeability field realizations. It is called STD in the remainder of this text. This overall standard deviation

for an EnKF method in a single synthetic experiment j is calculated as the square root of the mean over the domain of the pixel-wise ensemble variances, as follows

$$\mathrm{STD}_j = \sqrt{\frac{1}{n_g}\sum_{i=1}^{n_g}\sigma_{ij}^2}, \qquad j \in \{1, \cdots, 1000\}\,. \tag{5.16}$$

where n_g is the number of grid cells. The sample variances in Equation (5.16) are calculated as follows

$$\sigma_{ij}^2 = \frac{1}{n_e - 1}\sum_{k=1}^{n_e}\left(Y_{k,i,j} - \bar{Y}_{ij}\right)^2 \tag{5.17}$$

where n_e is the number of realizations in the ensemble, $\bar{Y}_{ij}$ is the mean logarithmic permeability for synthetic experiment j and grid cell i, and $Y_{k,i,j}$ is the permeability of realization k calculated at grid cell i in synthetic experiment j.

Besides the overall performance assessment combining results of the RMSE and STD, a smaller test with 10 synthetic experiments is carried out for checking the correlations between observed variables and the full parameter field. This input/output-intensive synthetic experiment is executed for the PP-EnKF and the classical EnKF. A reference data assimilation is computed using the EnKF with an ensemble size of 10 000. Subsequently, 10 synthetic experiments were computed for EnKF and PP-EnKF using ensemble size 50. For the tracer model, the correlations are of the form

$$\rho(c_{obs}, K_{\log}) = \frac{\mathrm{Cov}(c_{obs}, K_{\log})}{\sigma_{\mathtt{c}} \cdot \sigma_{\mathtt{K}}} \tag{5.18}$$

where the covariance Cov and the standard deviations $\sigma_{\mathtt{c}}$ and $\sigma_{\mathtt{K}}$ are estimated from the ensemble, c_{obs} $[N/L^3]$ denotes the concentration at one of the two observation locations, and $K_{\log}$ $[-]$ denotes the logarithmic permeability at a given location in the field. For the well model, the correlations are of the form

$$\rho(h_{obs}, K_{\log}) = \frac{\mathrm{Cov}(h_{obs}, K_{\log})}{\sigma_{\mathtt{h}} \cdot \sigma_{\mathtt{K}}} \tag{5.19}$$

where the difference to the tracer case is that h_{obs} $[L]$ denotes the head observed at one of the 49 measurement locations. By varying $K_{\log}$ across the model domain, I obtain the field of correlations for each observation location in a synthetic experiment. The RMSEs between these correlation fields and the reference correlation field is used to judge the amount of spurious correlation in the synthetic experiments introduced by the small ensemble size.

5.4 Results

5.4.1 Comparison of the PP-EnKF to other EnKF methods

First, results for the tracer model setup are shown. In Figure 5.1 (top), the mean RMSEs from 1 000 synthetic experiments are shown for the eight EnKF methods. The tracer model is known to yield a small range of RMSEs across the tested EnKF methods (see Chapter 4). Among the

eight tested EnKF methods, the PP-EnKF ranks among the four methods with the smallest RMSEs for all tested ensemble sizes. For ensemble size 50 and 250, the PP-EnKF has the fourth smallest RMSE. For ensemble sizes 70 and 100, it yields the third smallest RMSE.

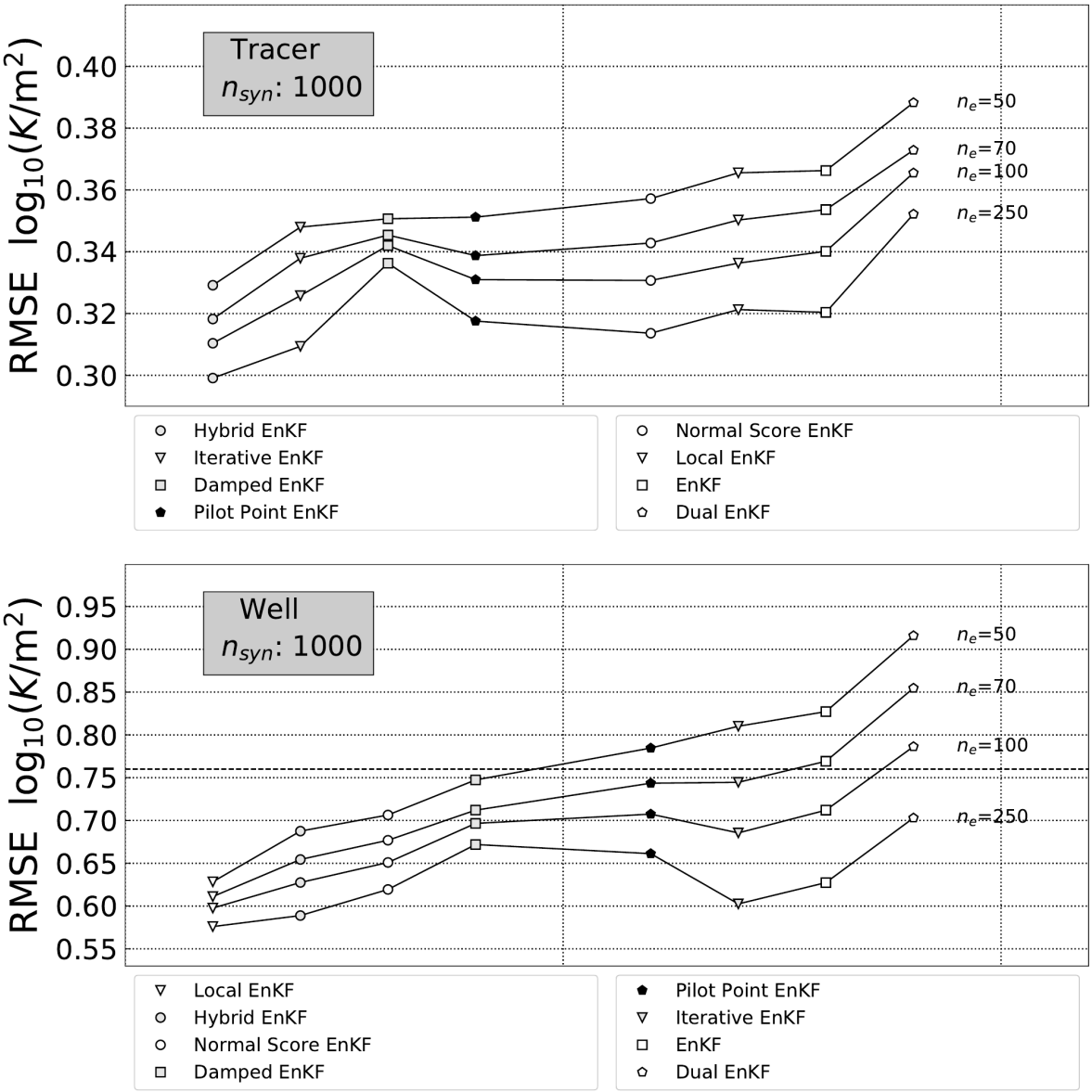

Figure 5.1: Comparison of the root-mean-square errors (RMSEs) of the pilot point EnKF and other EnKF methods in the tracer model (top) and well model (bottom).

A direct comparison to the classical EnKF is interesting for two reasons. First, the PP-EnKF is directly derived from the classical EnKF. Secondly, earlier publications include conflicting results regarding this comparison. In the synthetic experiments performed with the tracer model, the PP-EnKF performs significantly better than the classical EnKF for ensemble sizes 50 70, and 100, and slightly better for ensemble size 250.

Figure 5.1 (bottom) shows the RMSE means from the corresponding 1 000 synthetic experiments for the well model. Compared to the tracer model, the well model yields larger RMSE differences

between the tested methods. The PP-EnKF yields slightly worse than average RMSEs for the well model. For ensemble size 50 and 70, the PP-EnKF has the fifth smallest RMSE. For ensemble sizes 100, and 250, it scores the sixth smallest RMSE. Compared to the classical EnKF, the PP-EnKF performs better for ensemble sizes 50, 70, slightly better for ensemble size 100, and worse for ensemble size 250.

Summing up the results for the two physical model setups, the PP-EnKF yields good RMSE results for the tracer model and medium RMSE results for the well model. It should be noted that these setups were taken from the previous study in Chapter 4, and that a very robust testing was done using 1 000 synthetic studies. Nevertheless, the ranking of the EnKF variants differed considerably between the two model setups. It is therefore very well possible that, for other model setups, the PP-EnKF could outperform all other EnKF variants for small ensemble sizes. Results from both physical model setups suggest that, for small ensemble sizes, the PP-EnKF is a clear improvement to the classical EnKF as desired by design. This is in slight contradiction to results from Heidari et al. (2013), where the PP-EnKF yields larger RMSEs than the classical EnKF for a synthetic experiment with ensemble size 50. Our results for ensemble size 250, for which both methods yield similar results, are in agreement with similar results from Tavakoli et al. (2013).

Furthermore, the RMSEs suggest that the PP-EnKF provides a trade-off between damping and the classical EnKF. For ensemble size 50, the PP-EnKF has similar or larger RMSE than the damped EnKF, but smaller RMSE than the classical EnKF. Moving to larger ensemble sizes, the reduction of RMSE by the PP-EnKF is much larger than the reduction of RMSE by the damped EnKF. A possible explanation for this effect is that the interpolated updates of the PP-EnKF reduce spurious correlations and thereby reduce the number of divergent synthetic experiments for small ensemble sizes. For larger ensemble sizes, the relatively large reductions of the RMSEs of the PP-EnKF suggest that the interpolated updates of the PP-EnKF can incorporate more information than the damped updates of the damped EnKF.

The full RMSE distributions give a more in-depth picture of the performance of the EnKF methods. They are provided in Figures 5.2 and 5.3. The distributions illustrate how the RMSE means in the previous section were obtained. For the well model, one can see that the PP-EnKF yields a somewhat wider RMSE distribution than other methods.

5.4.2 Ensemble variance of conditional realizations

Now, I test and compare the uncertainty characterization of conditional realizations obtained from the PP-EnKF by looking at the overall ensemble standard deviations. The equations for the overall standard deviations were introduced in Section 5.3.4. I compare the mean overall standard deviations calculated over 1 000 synthetic experiments for each EnKF method. Additionally, a benchmark STD is derived from 100 synthetic experiments using the classical EnKF with ensemble size 10 000.

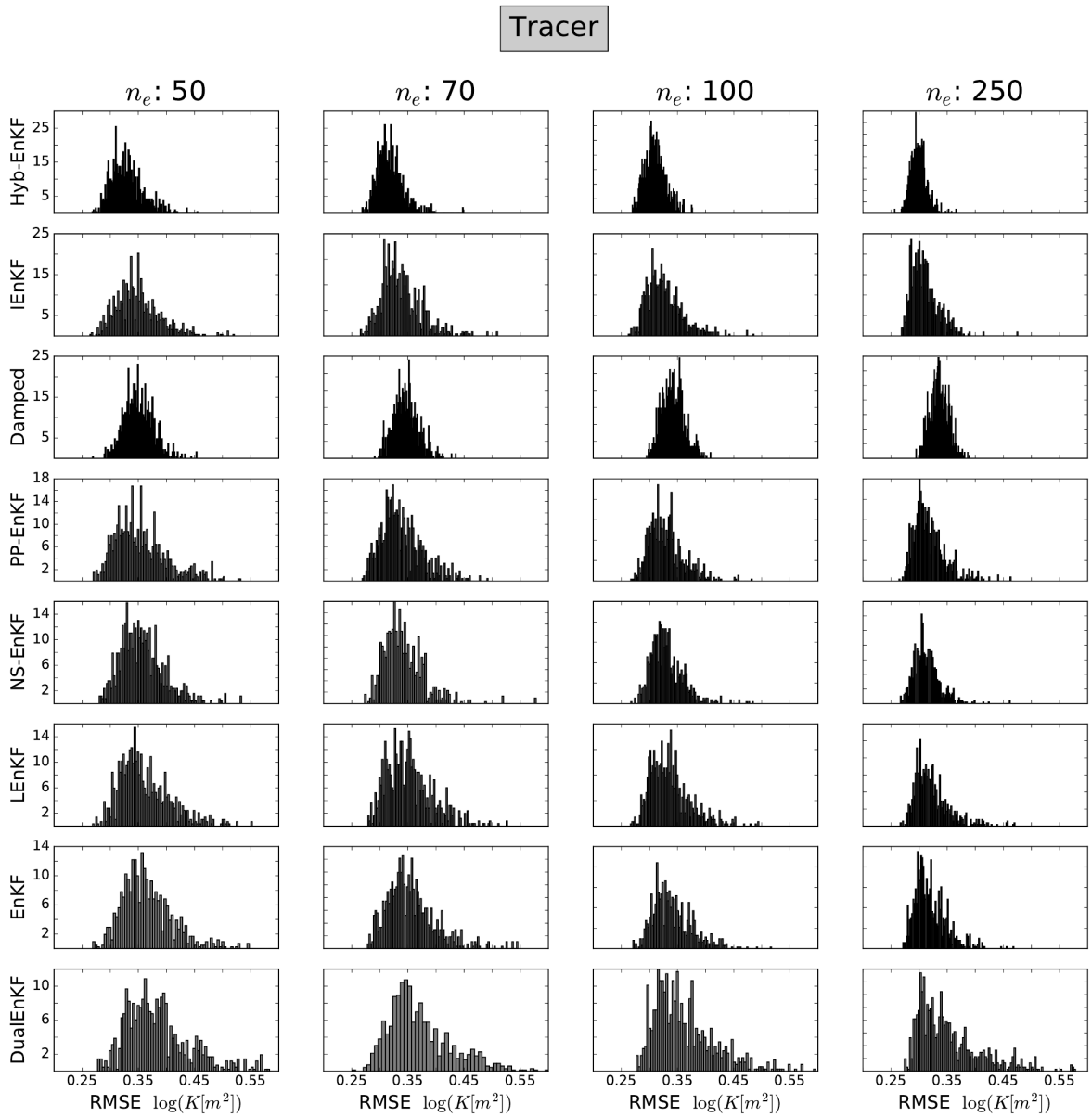

Figure 5.2: Root-mean-square error distributions of pilot point EnKF and other EnKF methods in the tracer model. The order of methods is taken from the RMSE results for the tracer model from Figure 5.1.

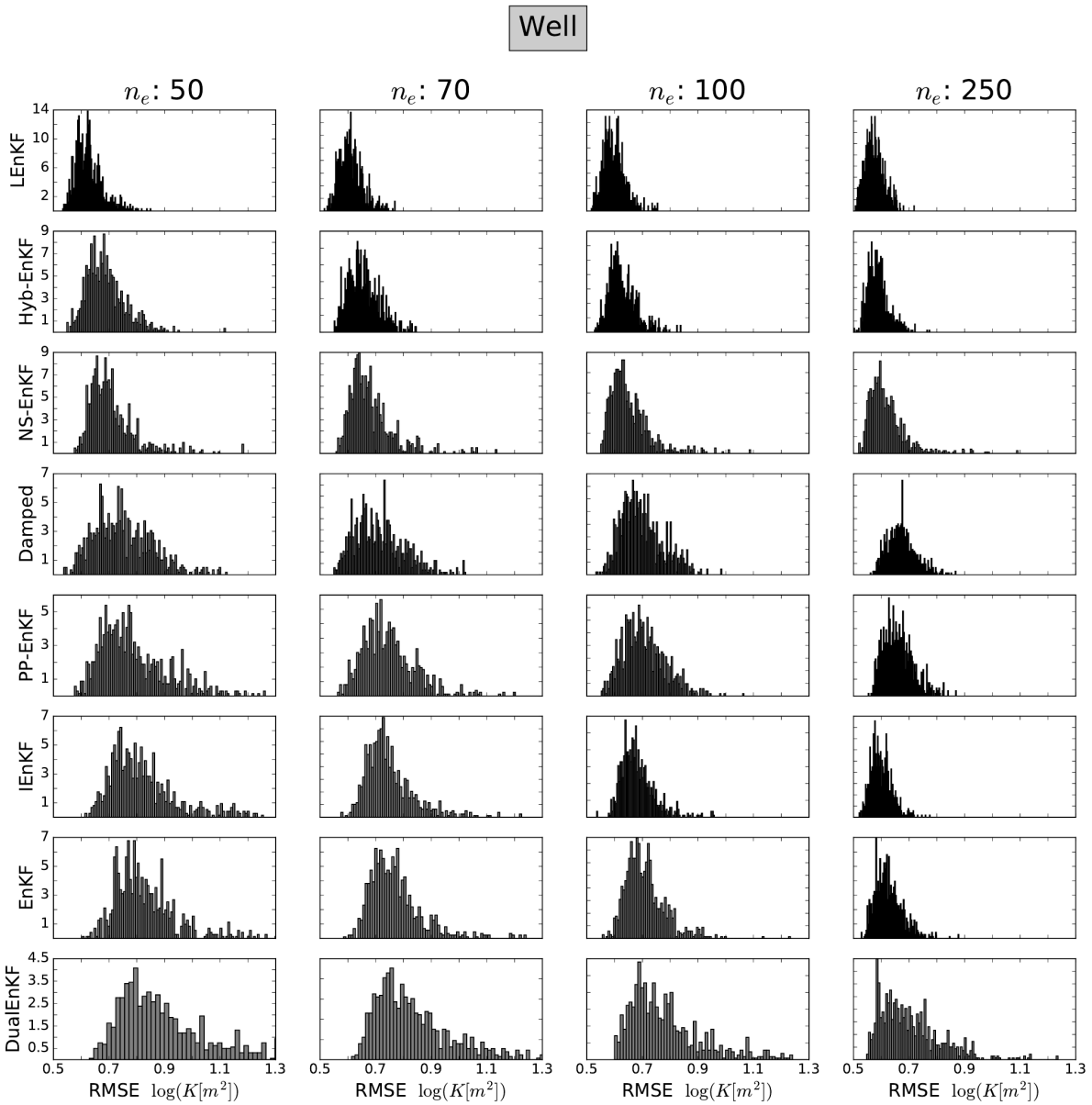

Figure 5.3: Root-mean-square error distributions of pilot point EnKF and other EnKF methods in the well model. The order of methods is taken from the RMSE results for the well model from Figure 5.1.

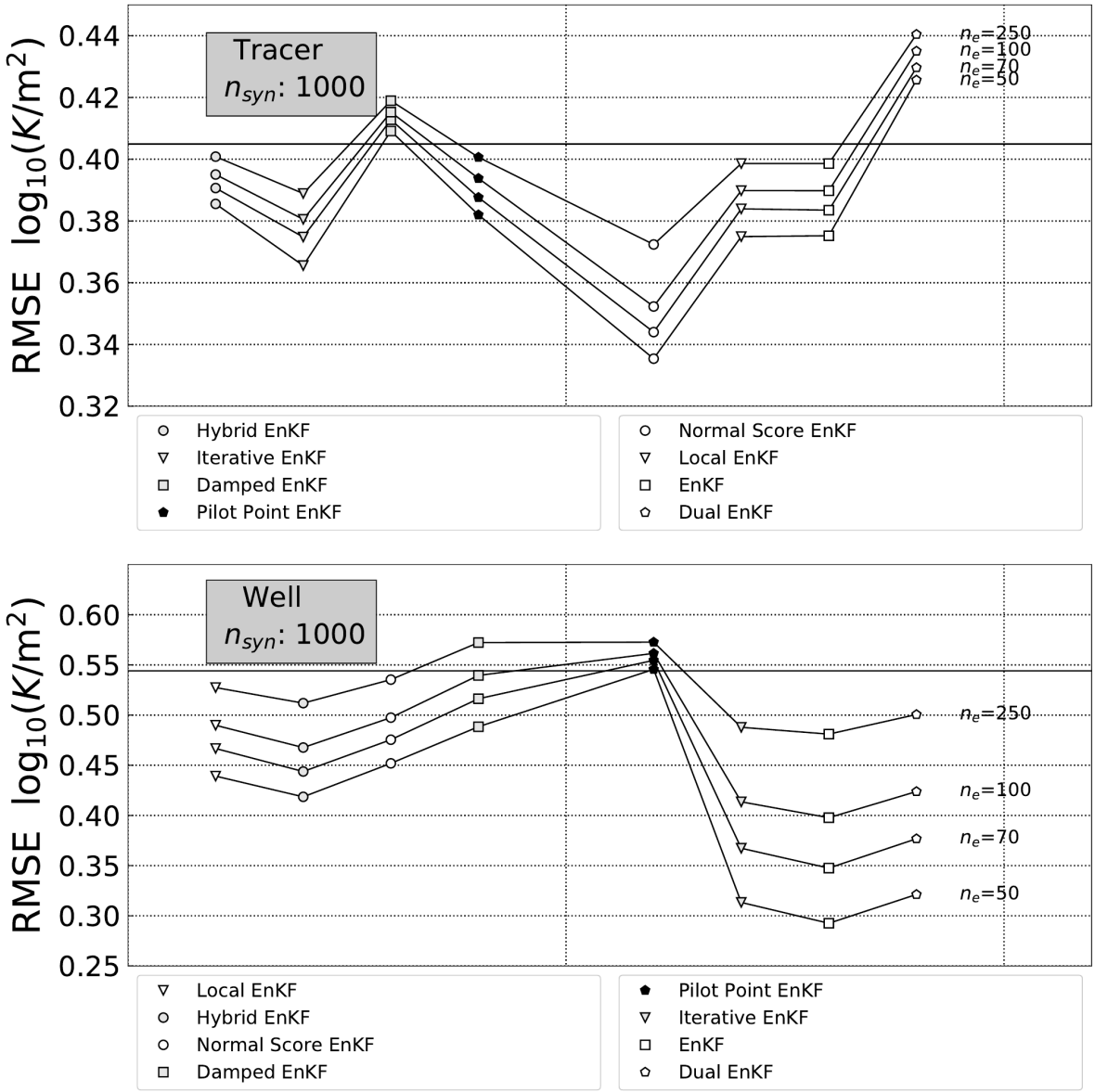

Figure 5.4: Comparison of the overall standard deviation (STD) of the pilot point EnKF and other EnKF variants for the tracer model (top) and well model (bottom). The horizontal black line depicts the STD obtained from 100 synthetic experiments with 10 000 ensemble members.

The results for the tracer model are shown in the top half of Figure 5.4. All EnKF methods yield standard deviations that are reasonably close to the standard deviation of the benchmark run. Still, most methods underestimate the ensemble spread; exceptions are the damped EnKF and the dual EnKF. Comparing the differences to the 10 000 ensemble run, the pilot point EnKF ranks fourth for ensemble size 50, third for ensemble sizes 70 and 100, and second for ensemble size 250. Thus, the pilot point EnKF is in the top half of the methods, not only for RMSE comparison, but also concerning the standard deviation.

For the well model, the mean overall standard deviations are shown in the bottom half of Figure 5.4. Compared to the tracer model, there is a larger tendency of most methods to underestimate ensemble spread compared to the 10 000 ensemble run. Additionally, one can see a clear difference between the EnKF methods that contain a form of damping and the ones that do not (including the iterative, the classical and the dual EnKF). The pilot point EnKF yields a large ensemble spread, especially for ensemble size 50, where it yields the best STD of all methods. For ensemble size 70, the pilot point EnKF still ranks first, for ensemble size 100 it ranks second and for ensemble size 250 it ranks third. Thus, while the pilot point EnKF only obtained medium results in RMSE comparison compared to other EnKF methods, it always ranks in the top three methods concerning uncertainty characterization, ranking first for the important smallest ensemble sizes.

In conclusion, the uncertainty characterization of the pilot point EnKF is generally good compared to other EnKF methods. This is one of the main features of the method, and this benefit again comes by design. In the update step, the faulty reduction of ensemble variance is constrained by two effects. First, by reducing the number of parameters to the number of parameters at pilot points, issues of rank deficiency and inbreeding are reduced, and secondly, by interpolating the update, large parts of the prior variability remain intact. These results agree with results from Heidari et al. (2013), where a large spatial variability in PP-EnKF updates is diagnosed. While Tavakoli et al. (2013) also diagnose heterogeneities in the PP-EnKF results, the small ensemble spread in their PP-EnKF results is contradictory to the findings from this work. This might be due to the specific subsurface model in Tavakoli et al. (2013), where only a small fraction of the model parameters was updated.

5.4.3 Variation of prior correlation lengths

In this section, results for faulty correlation lengths are discussed: the prior correlation lengths are varied to half or twice the correlation length of the synthetic truth. These results are especially important for judging the performance of the PP-EnKF, since it is influenced by the correlation length through both the prior realizations and the kriging interpolation.

First, I discuss the tracer model with correlation length 25 m. This correlation length is half the correlation length of the synthetic truth. Figure 5.5 (top) shows the average RMSE values comparing to the results in Section 5.4.1. For ensemble size 50, all EnKF methods yield larger RMSEs than for the standard case. For ensemble size 250, all methods except the classical

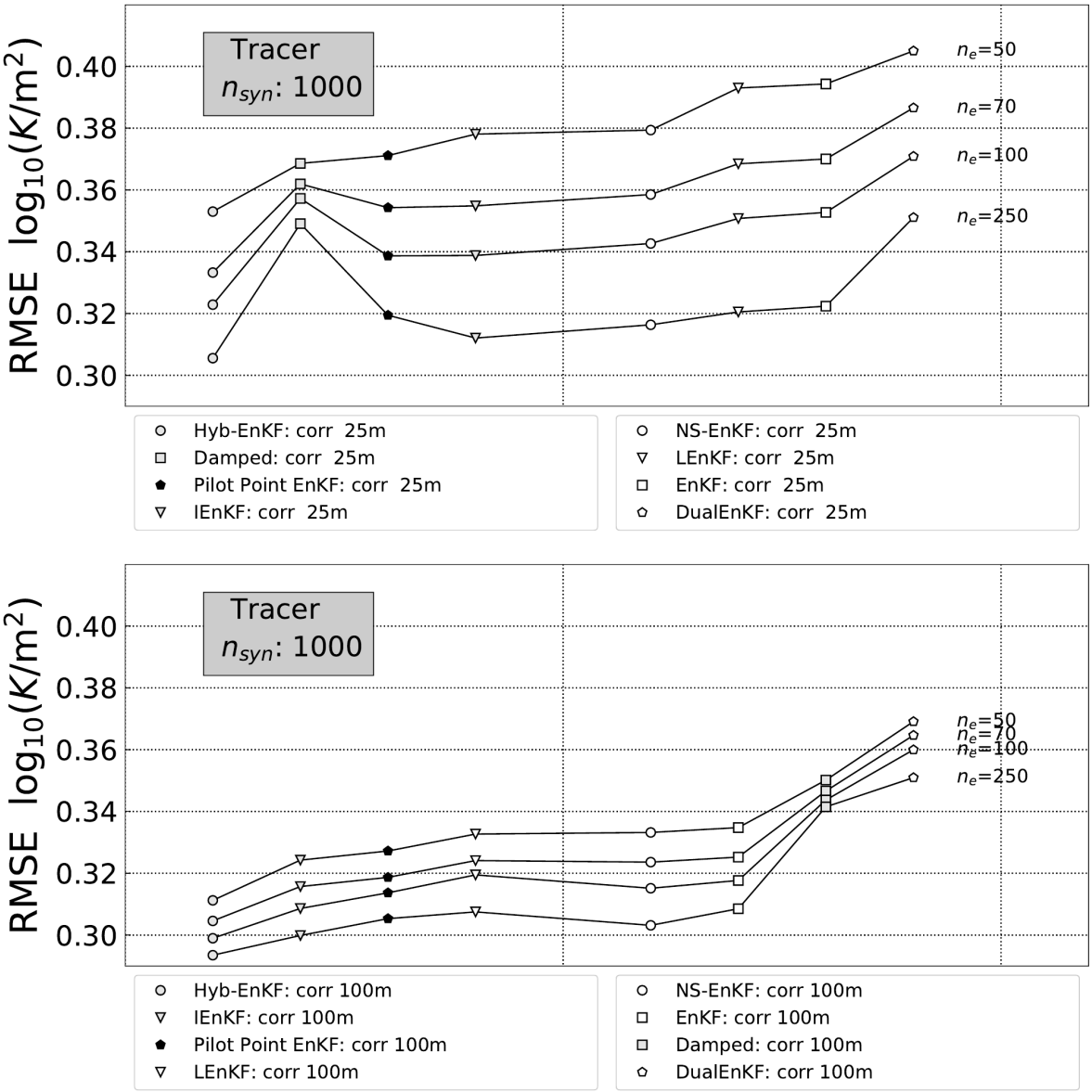

Figure 5.5: Comparison of the pilot point EnKF with other EnKF variants for the tracer model and correlation lengths 25 m (top) and 100 m (bottom).

EnKF and the dual EnKF yield larger RMSEs than for the standard case. A specific look at the PP-EnKF shows that, for ensemble size 50, it ranks third among the EnKF methods. For ensemble sizes 70 and 100, it ranks secondly, and for ensemble size 250, third. Relative to the other EnKF methods, these results are a slight improvement for the PP-EnKF compared to the case of the correct correlation length. This means that the PP-EnKF seems to be robust against a too small specification of correlation length for this test case.

Now I discuss the results for the tracer model and a too long correlation length of 100 m as shown in Figure 5.5 (bottom). For all ensemble sizes, all methods except the damped EnKF yield smaller RMSEs than for the correct correlation length (50 m). The damped EnKF yields slightly larger RMSEs for ensemble sizes 100 and 250 than for the standard case. Regarding the PP-EnKF, it ranks third among all methods for ensemble sizes 50, 70 and 100. For ensemble size 250, it ranks fourth among all methods. Thus, the ranking of the PP-EnKF among the different methods is comparable to the corresponding results for the correct correlation length.

For a discussion of these results, recall how a faulty correlation length may affect the performance of EnKF methods. The correlation length primarily affects the prior permeability fields. This should generally be a disadvantage for the update, thus leading to larger RMSEs. An additional effect of the correlation length is its direct influence on the EnKF update. The correlations that drive the update are either more restricted to the vicinity of the measurement locations, for the case of a small correlation length, or they are spread out more widely for the case of the large correlation length. The RMSE results for the small correlation length suggest that the restriction of the EnKF update is particularly obstructive for the tracer model, possibly because there are only two measurement locations in this setup. Thus, for a small number of measurement locations, an underestimation of the correlation length may significantly inhibit EnKF updates. On the other hand, the smaller RMSEs for correlation length 100 m are surprising. This result suggests that the effect of the larger update radius is stronger than the effect of the wrong prior correlation length in the tracer setup. Good results for too large correlation lengths have been documented in the literature, for example by Chaudhuri et al. (2018). This effect may vanish for more complicated synthetic reference fields, especially, when they exhibit small-scale heterogeneities.

Turning to the well model, the PP-EnKF for correlation length 30 m is tested. Figure 5.6 (top) shows RMSEs calculated over 1 000 synthetic experiments. Again, for all ensemble sizes, all EnKF methods yield slightly larger RMSEs than for the case of the correct correlation length (60 m). The PP-EnKF ranks fifth among the EnKF methods for ensemble sizes 50 and 70. For ensemble size 70 it ranks sixth and for ensemble size 250 seventh. This almost reproduces the results for the correct correlation length, only for ensemble size 250 the rank is slightly worse.

I now look at results for the well model and correlation length 120 m in Figure 5.6 (bottom). Here, for all ensemble sizes, all EnKF methods yield slightly smaller RMSEs than for the standard case. Regarding the PP-EnKF, it ranks sixth for ensemble sizes 50, 70 and 250, and it ranks seventh for ensemble size 100. Thus, again the ranking of the PP-EnKF among the methods is slightly worse than for the correct correlation length.

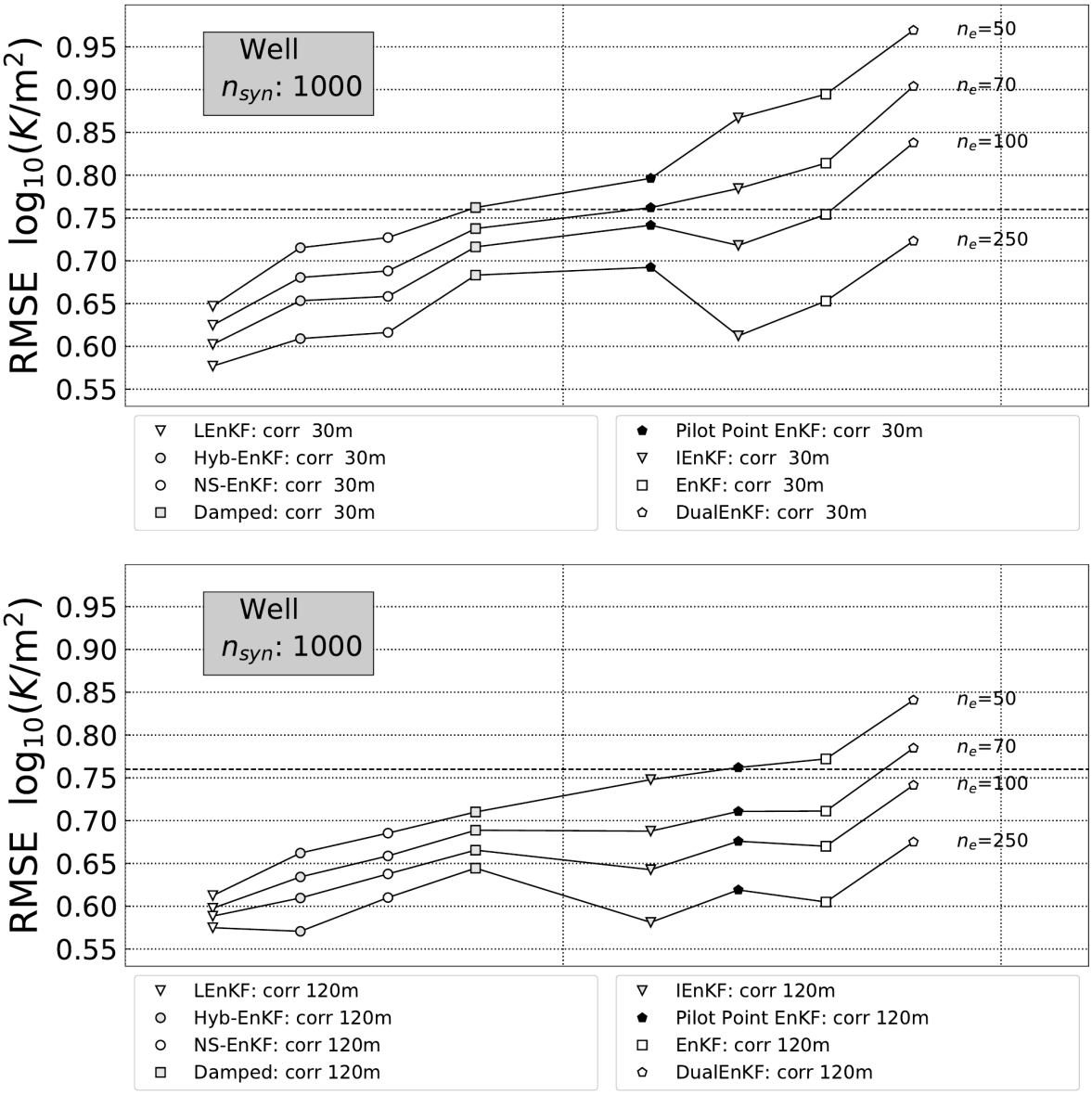

Figure 5.6: Comparison of the pilot point EnKF with other EnKF variants for the well model and correlation lengths 30 m (top) and 120 m (bottom).

In summary, the PP-EnKF is not affected more by a misspecification of the correlation length than other EnKF variants for both model setups, even though the correlation length has a more pronounced role in the PP-EnKF than in other variants.

5.4.4 Average rankings of the PP-EnKF

In the last sections, various types of synthetic experiments were used to compare the performance of the PP-EnKF to other EnKF methods. This included synthetic experiments with the correct and erroneous prior correlation lengths, and repetitions for ensemble sizes of 50, 70, 100 and 250. In Table 5.1, the average RMSE rankings are displayed that are calculated from the twelve single RMSE rankings (four ensemble sizes for each of three correlation lengths). For the tracer setup, the PP-EnKF has the third best average ranking. Thus, for the tracer setup, the PP-EnKF has very good average results outperformed only by the hybrid EnKF and the iterative EnKF. For this setup, the performance of the PP-EnKF justifies its usage over the majority of other EnKF methods. For the well setup, the PP-EnKF only has the sixth best average ranking. For this setup, the average RMSE does not justify using the PP-EnKF compared to existing EnKF variants.

Table 5.1: Average RMSE rankings of the eight EnKF methods for tracer and well setups. The average ranking is the mean of the twelve rankings from the RMSE comparisons of all tested prior correlation lengths.

RMSE	Tracer	Well
EnKF	6.1667	6.3333
Damped	5.9167	4.9167
NS-EnKF	3.9167	3.3333
DualEnKF	8.0	8.0
Hyb-EnKF	1.0	1.9167
LEnKF	5.4167	1.0833
IEnKF	2.3333	4.5833
PP-EnKF	3.25	5.833

Table 5.2: Average STD rankings of the eight EnKF methods for the tracer and well setups. The average ranking is the mean of the four rankings from the overall STD comparison.

STD	Tracer	Well
EnKF	4.75	8.0
Damped	2.25	2.0
NS-EnKF	7.75	2.5
DualEnKF	6.0	6.0
Hyb-EnKF	1.5	5.0
LEnKF	4.25	3.5
IEnKF	6.5	7.0
PP-EnKF	3.0	2.0

The reproduction of the correct ensemble variance is another key performance measure for comparing the EnKF methods. Therefore, overall STD results for the four ensemble sizes 50, 70,

100, 250 were ranked and subsequently averaged, and results are summarized in Table 5.2. The PP-EnKF performs well among the eight tested methods. For the tracer setup, the PP-EnKF ranks third out of eight methods. In particular, the PP-EnKF ranks significantly better than the iterative EnKF that ranked better for average RMSE. Thus, taking into account both, average RMSE and average STD, the PP-EnKF is outperformed only by the hybrid EnKF. For the well setup, the PP-EnKF ranks first together with the damped EnKF. For the well setup and for small ensemble sizes, the PP-EnKF ranks uniquely first.

5.4.5 Variation of pilot point grids

One important degree of freedom added by the PP-EnKF is the choice of the pilot points. In this section, RMSEs for different pilot point grids are compared in the tracer and well model.

For the tracer model, the compared grids are regular square grids with different numbers of pilot points. All grids also include the two measurement locations of the tracer model. The results in Figure 5.7 (top) show small changes of the RMSE for different grids, all changes are smaller than 0.01.

For the well model, the compared grids are extensions of the standard grid, since the standard grid is made up of all measurement locations and, by the definition of the PP-EnKF, I want to keep the measurement locations as pilot points in all grids. The first extension is adding pilot points on the diagonal between the original pilot points. The second extension is another regular grid with twice the number of pilot points in each column and row. Again, the results in Figure 5.7 (bottom) exhibit small changes of the RMSE, only the doubled grid yields significantly smaller RMSE than the standard method.

The results of the grid variation have two implications. First, the differences between the grids are not very large, thus at least in the two setups treated here, the PP-EnKF shows a certain robustness against different choices of regular grids in the investigated range. Secondly, the results do also indicate that a smart choice of grid might make a bigger difference in other setups. While the grid with the smallest number of pilot points yields the smallest RMSEs for the relatively homogeneous synthetic true permeability field of the tracer model, the grid with the largest number of pilot points yields the smallest RMSEs for the more heterogeneous well model. This suggests a model-specific number of pilot points required to sufficiently approximate the real cross-correlation functions. As a general rule, I refer to the recommendations for pilot point spacing in the literature (Capilla et al., 1997) that suggest about three pilot points per range as a robust choice. The range from Capilla et al. (1997) would correspond to two correlation lengths in this study, a distance at which parameters are barely correlated.

5.4.6 Spurious correlation reduction

Now the ability of the PP-EnKF to reduce spurious correlations is tested by checking the correlations between observed variables and updated permeabilities in the tracer and well setup

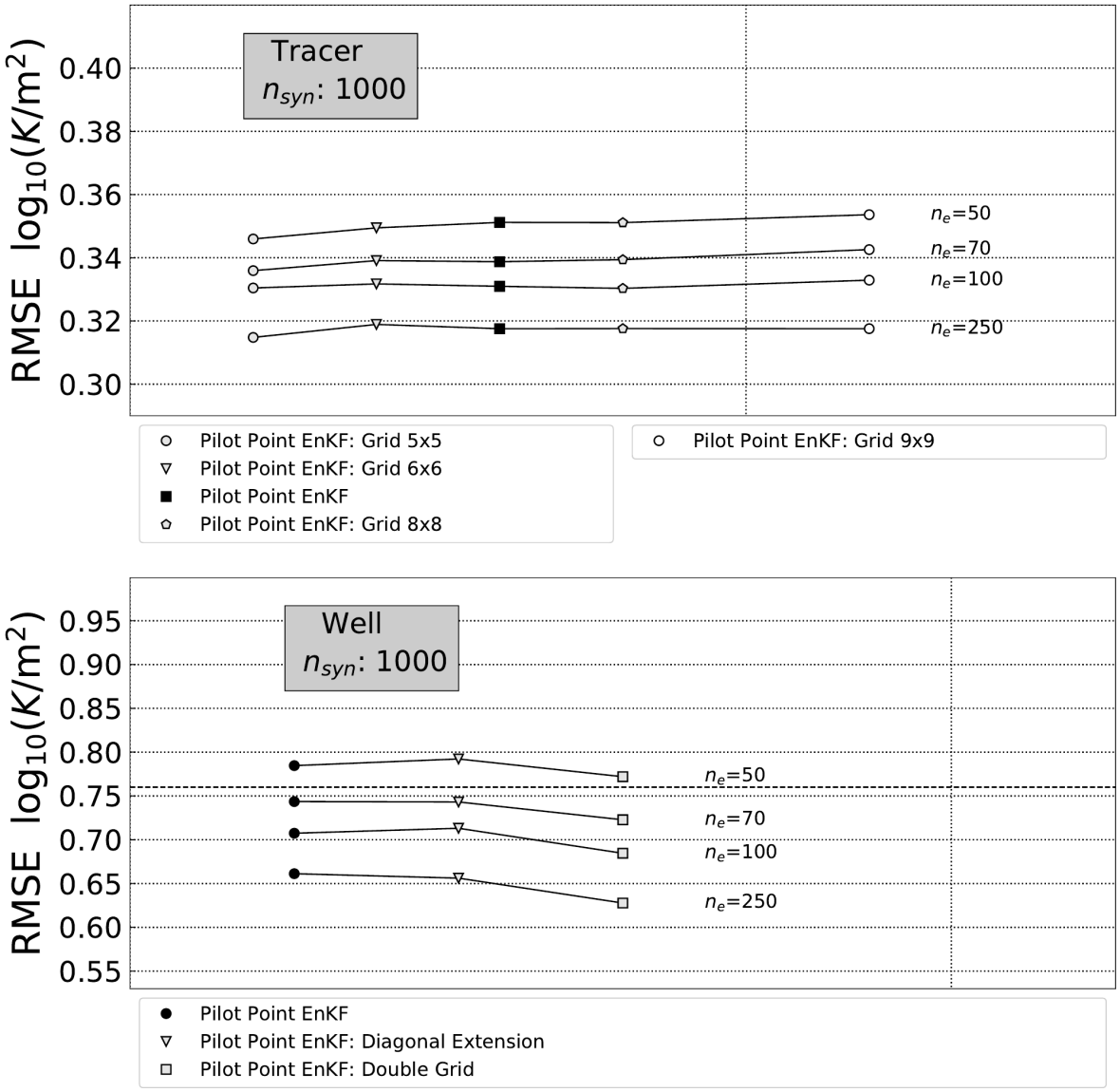

Figure 5.7: Mean RMSE (calculated over 1 000 synthetic experiments) for PP-EnKF with different regular grids for the tracer setup (top) and well setup (bottom).

after assimilation. As a benchmark, I use a reference synthetic experiment with the EnKF and 10 000 ensemble members. Figure 5.8 shows the reference correlation fields for the tracer and well setup after the EnKF updates. For the tracer setup, both observation locations are shown. The concentrations at the observation location in the highly permeable region (compare the synthetic true permeability fields from Figure 4.1) yields comparably small correlations. The observation location in the low-permeable region yields large positive correlations with the low-permeability region south of the observation location. Turning to the well setup, two representative head observation locations are displayed. One can observe a general trend, small positive correlation in the direction of the injection well in the center of the model and large negative correlation in the direction of the model boundaries. Note that the correlation structure in the well setup is finer and more pronounced than in the tracer setup. This is related partly to the characteristics of the permeability fields.

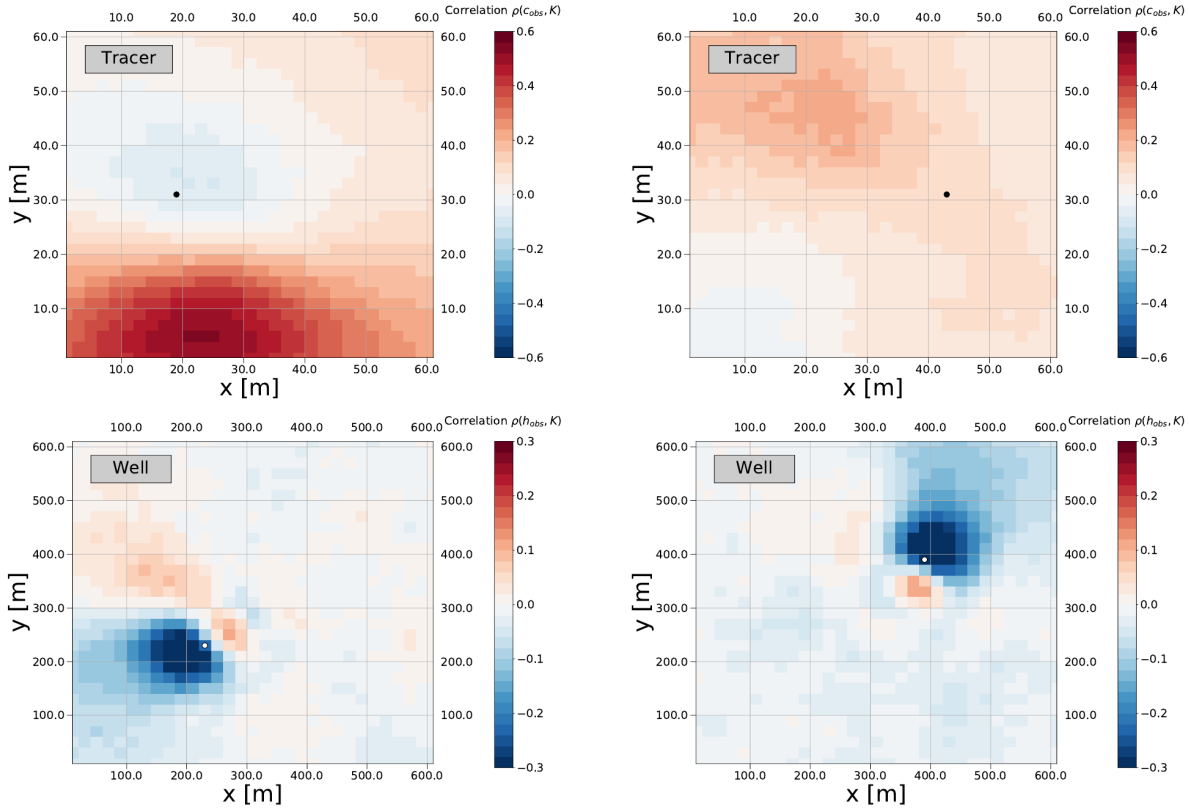

Figure 5.8: Reference correlations between an observed variable at a chosen observation location and the permeability throughout the model domain. For the tracer setup (top), the two observation locations for concentration are plotted in black, for the well setup (bottom) the two (out of 49) observation points for head are plotted in white.

Now I compare correlation fields from 10 synthetic experiments (for the PP-EnKF and the classical EnKF using ensemble size 50) to the correlation fields of the reference. Table 5.3 shows the RMSE values for each synthetic experiment. For the tracer setup in five synthetic experiments, PP-EnKF is closer to the reference than for EnKF, and for five synthetic experiments the opposite holds. The mean RMSEs for the PP-EnKF and the EnKF for the ten synthetic experiments

Table 5.3: Root-mean-square errors of correlation fields (compared to reference correlation fields).

	RMSEs					Mean
Tracer (EnKF)	0.133	0.167	0.215	0.139	0.154	0.161
	0.163	0.266	0.106	0.125	0.141	
Tracer (PP-EnKF)	0.142	0.169	0.213	0.128	0.155	0.162
	0.161	0.260	0.113	0.137	0.137	
Well (EnKF)	0.180	0.159	0.168	0.172	0.178	0.173
	0.173	0.171	0.190	0.162	0.173	
Well (PP-EnKF)	0.152	0.153	0.145	0.149	0.158	0.149
	0.142	0.149	0.147	0.145	0.147	

are very close. For the well setup, PP-EnKF has smaller RMSE values than the EnKF for all ten synthetic cases, with on average a smaller RMSE of 13.9 %.

The results from this section support results from the overall standard deviation STD from Figure 5.4. For the tracer setup, the similar results from this section correspond to relatively similar overall STDs of the EnKF and the PP-EnKF in Figure 5.4. For the well setup, the PP-EnKF resulted in a larger STD compared to the EnKF. In this section, I additionally find a better characterization of the spatial correlation structure by the PP-EnKF, compared to the EnKF. Thus, the larger overall uncertainty shown by the STD does not stem from additional noise, it is rather the result of a correlation structure closer to a large-ensemble run. To summarize results for the synthetic experiments for ensemble size 50, the PP-EnKF has both a better uncertainty characterization and a better RMSE than the EnKF. This illustrates the significant advantages of the PP-EnKF compared to the EnKF.

5.5 Conclusion

The ensemble Kalman filter is a powerful tool for parameter estimation used in the geosciences. This work introduces the PP-EnKF designed to reduce spurious covariances for small ensemble sizes. The parameter field that is estimated is split up into pilot points and non-pilot points. The EnKF update is restricted to parameters at pilot point locations and to dynamic variables. After the update, a kriging interpolation of the update is applied to update parameter values at non-pilot point positions. The initially seeded spatial variability in between the pilot points remains untouched. The implementation of the PP-EnKF is relatively straightforward, since the definition of the state vector and the kriging interpolation can be implemented as pre- and post-processing steps around the implementation of the classical EnKF update. The stabilizing effect of the PP-EnKF can be compared with the hybrid EnKF and the local EnKF. I claim that the main theoretical advantage of the PP-EnKF compared to these methods is that its update at pilot points is computed from the unmodified ensemble covariance matrix. In the local EnKF and the hybrid EnKF, the covariance matrix is explicitly modified. The PP-EnKF introduces two additional input parameters compared to other EnKF methods, the locations of

the pilot points, and the covariance matrix of the kriging interpolation. It is suggested to choose pilot points on a regular subgrid of the model domain. Additionally, all measurement locations should be included as pilot point locations. In general, one could vary the density of pilot point locations according to given prior information. In regions, where large updates are expected, the number of pilot points can be increased. The kriging covariance matrix of the PP-EnKF is defined according to the prior variogram model, which is also used as input to generate random fields for all EnKF methods.

The PP-EnKF compares well to other EnKF methods for two physical model setups, a solute transport model and a model around an injection well. This is concluded from an extensive comparison of RMSEs between parameter estimation results from 1 000 synthetic experiments and a synthetic true parameter field. Even for synthetic experiments with erroneous prior correlation lengths (half and twice the correct correlation length), the performance of the PP-EnKF remains similar to the case with the correct prior correlation length. This is important, as the prior correlation length plays a special role in the PP-EnKF method. In the average RMSE ranking, the PP-EnKF is third best for the tracer and sixth best for the well setup. Compared to the EnKF, the PP-EnKF is performing significantly better in both setups.

The PP-EnKF ranks particularly well against the other EnKF methods regarding the preservation of the spatial variability throughout the estimation. Especially, for ensembles sizes of 50 and 70 and for the well model, the PP-EnKF yields the ensemble variance that compares best to a test run with the EnKF and a very large ensemble size of 10 000. In an average overall STD ranking, the PP-EnKF ranks third best for the tracer setup and best in the well setup. Additionally, distributed correlation fields of the PP-EnKF and the classical EnKF were compared. For the well setup, the correlations of the PP-EnKF are significantly closer to a reference field than the correlations of the classical EnKF. For the tracer setup, the correlations of the PP-EnKF and the classical EnKF are equally close to the reference. Reproducing the posterior variance is an important feature of an EnKF method, since many EnKF methods suffer from an underestimation of the posterior variance that may lead to wrong interpretation of results or in the worst case to filter divergence. The PP-EnKF not only ranks particularly well against the other EnKF methods regarding the reproduction of the posterior variance. For a small ensemble, it also reproduces spatially distributed correlation fields better than the classical EnKF. This suggests that the PP-EnKF is able to preserve better than the EnKF not only the variance in the permeability field, but also the correlations of the permeability field with the dynamic variables.

From the aforementioned discussions I conclude that the PP-EnKF outperforms clearly the standard EnKF, and outperforms most EnKF variants for the reproduction of the ensemble spread. It is therefore a very interesting EnKF variant, with the need for further research to investigate issues like the optimal placing of pilot points and other applications like the estimation of soil hydraulic parameters.

Aus der Wahrheit Feuerspiegel
lächelt `sie` *den Forscher an.*

An die Freude
Friedrich Schiller

6 Conclusion

The overall objective of this thesis is to present a framework for implementing and robustly comparing ensemble Kalman filter (EnKF) methods. These EnKF methods are applied for parameter estimation in groundwater flow and transport problems. The numerical software SHEMAT-Suite is presented, in particular the stochastic mode of SHEMAT-Suite that contains the implementation of the EnKF methods from this thesis. Then, a number of EnKF methods is robustly compared and a new EnKF method, the pilot point ensemble Kalman filter (PP-EnKF), is introduced. The following three main conclusions can be drawn from this thesis:

- Conclusion 1: The implementation of the suite of EnKF methods in the numerical software package SHEMAT-Suite lead to the formulation of the following three guidelines for scientific programming: (1) open-source code, (2) modularity, (3) a distributed workflow using git branches.
- Conclusion 2: When EnKF methods are used for estimating parameters with small ensemble sizes, the uncertainty coming from the choice of the random seed is non-negligible. Consequently, multiple (at least 10) parameter estimation experiments are needed for assessing the relative performance of these EnKF methods.
- Conclusion 3: The PP-EnKF is an EnKF method useful for parameter estimation. The PP-EnKF reduces two problems associated with parameter estimation using EnKF methods: spurious correlations and an underestimation of posterior variances.

In the following, I present these conclusions in more detail alongside possible future perspectives.

Chapter 3 I give an introduction to the software SHEMAT-Suite, a parallelized open-source code for simulating flow, heat and mass transport in porous media. A suite of EnKF methods was implemented in the so-called stochastic mode of SHEMAT-Suite. Three topics from this implementation are now highlighted: (1) open-source software, (2) modular implementation, and

(3) a workflow using the version control software git. First, open-source software is important for the scientific process, especially concerning reproducibility and the ability to check results. For SHEMAT-Suite, the workflow for software development is divided between open-source published versions and internal code development. On the one hand, the major versions of SHEMAT-Suite are published under the MIT license. On the other hand, small-step scientific software development is carried out at research institutes as closed-source. The second topic, modularity, is another important guideline for the software development of SHEMAT-Suite. Modularity is applied for various parts of SHEMAT-Suite. For example, the source code of the non-linear iteration of the time-stepping is separated from the implementation of the linear solver. This facilitates the implementation of alternative linear solver software. Another example of modularity in SHEMAT-Suite are the property modules that define the functional relationships of subsurface parameters. These property modules are a crucial user-interface for setting subsurface parameters, for example based on laboratory experiments. Therefore, the property modules are separated from the rest of the code in order to facilitate user interaction. Concerning the stochastic mode of SHEMAT-Suite, the implementation of EnKF methods is separated from the EnKF update whenever this is possible. This facilitates the implementation of new EnKF methods, as well as the comparison of existing EnKF methods, because the EnKF methods use as much common computation steps as possible. Finally, the parameter estimation and inversion capabilities of SHEMAT-Suite are separated from the pure forward computation. This way, a fast and simple forward code is maintained, at the same time implementing the more sophisticated inversion techniques. Separating the parameter estimation and inversion capabilities of SHEMAT-Suite from the pure forward computation is connected to the third topic: a workflow for using the version control software git for developing SHEMAT-Suite. In SHEMAT-Suite, git branches are used most prominently for developing the three main source code versions, called modes: (1) forward mode, (2) stochastic mode, and (3) AD-mode for inversion using automatic differentiation. Additionally, git branches are used for implementing performance enhancements such as parallelization and automatic differentiation. These performance enhancements require additional source code and employing git branches It is possible to keep the core part of the source code simple, while still using the advantages of the performance enhancements in the specialized branches. Finally, git branches are used in their classical usage for implementing additional functionalities of SHEMAT-Suite.

Outlook: A number of developments can be implemented in SHEMAT-Suite building on the work presented in this thesis. Regarding further software development, the testing of SHEMAT-Suite could be enhanced in one of the three following ways. First, the existing suite of test models for output comparison should be continually extended, and more realistic test simulations could be introduced to enhance the testing. Additionally, the testing of SHEMAT-Suite could be automated and more closely linked to the software development of SHEMAT-Suite by following a continuous integration workflow. Another step for enhancing the testing of SHEMAT-Suite would be implementing unit tests for specific parts of the software, such as the linear solvers or the EnKF update equation. Aside from the testing, it is planned to use the modularity of SHEMAT-Suite for linking SHEMAT-Suite to other available software, for example for including advanced input and output handling, or parallelized data assimilation algorithms. Moreover, the

modular software development makes additional performance studies possible. For example, one can compare the pure forward computations of SHEMAT-Suite to another computation using the parallelized code version or the code with automatic differentiation. Finally, the git workflow that was applied to SHEMAT-Suite is easily generalized. For avoiding trade-offs between readable and fast source code, I propose adopting this workflow in other scientific software development projects.

Chapter 4 From the performance comparison of EnKF methods in Chapter 4 it can be concluded that multiple parameter estimation experiments are needed for assessing the performance of EnKF methods. I compare EnKF methods on the basis of a comparison between the root mean square error (RMSE) for the updated parameter field and a synthetic true parameter field. In particular, the differences between these RMSE scores for different EnKF methods are compared. The question of the significance of these relative RMSE differences arises. To clarify this, I derive thresholds on relative RMSE differences that can be significantly detected using 1, 10, or 100 synthetic parameter estimation experiments. Single synthetic experiments are generally not enough to show a significant difference in performance of EnKF variants, even when these EnKF variants result in RMSE differences of 15 %. In this study, even 10 synthetic experiments were not enough to show that relative RMSE differences between EnKF variants smaller than 10 % are significant. On the other hand, in this work and for two different groundwater flow model setups, 100 synthetic experiments allow to show that RMSE differences between EnKF variants larger than 2 % are significant. As in addition other sources of uncertainty play a role in real-world studies and could be considered as well in comparison studies, I conclude that at least 10 synthetic experiments are needed to rigorously compare two EnKF variants.

Outlook: On the basis of these results, I propose carrying out a larger number of and more profound EnKF method comparisons. In particular, the results of the study from Chapter 4 could be extended by including additional methods in the comparison, for example including combinations of methods. Recently introduced candidates for additional method comparisons include forward model simplification methods, such as principal component analysis and reduced basis methods, or ensemble smoother methods, like the iterative ensemble smoother. If the computational resources exist, the comparison should be extended to more and larger simulations, for example using three-dimensional simulations. In general, finding performance advantages of EnKF methods that are specific to certain types of simulations would be interesting. For example, the relative performance of an EnKF method could be linked to the relative heterogeneity of a parameter field. Another possible extension of the testing framework would be explicitly including the runtime of the methods. In the comparison in this thesis, I wanted to compare the performance independently of the runtime, but a more sophisticated performance measure could be developed that measures both, the update performance and the runtime. While a minimum number of necessary synthetic experiments was found in this thesis, a more in-depth analysis of the number of synthetic experiments needed for certain comparisons could be carried out. Finally, applying the robust comparison from this work to more realistic simulations and to data assimilation experiments using real observations can yield interesting insights in the performance

of EnKF methods.

Chapter 5 I introduce the PP-EnKF. The PP-EnKF is compared to other EnKF methods in two groundwater model flow setups: a tracer setup computing groundwater flow and solute transport, and a well setup computing groundwater flow around a central injection well. In the average RMSE ranking, the PP-EnKF is third best for the tracer setup and sixth best for the well setup. Compared to the EnKF, the PP-EnKF is performing significantly better in both setups. For the tracer setup the PP-EnKF is even better than most other enhanced EnKF methods. Additionally, I show that the PP-EnKF is stable with respect to the choice of prior correlation length and the choice of different pilot point grids. The preservation of the spatial variability throughout the estimation is a particular strength of the PP-EnKF. Especially, for comparably small ensemble sizes of 50 and 70 and for the well model, the PP-EnKF yields the ensemble variance that compares best to the reference ensemble variance (computed with a very large ensemble) out of all EnKF methods. In order to verify that the PP-EnKF indeed suppressed spurious correlations, distributed correlation fields computed by the PP-EnKF and the classical EnKF were compared. For the well setup, the correlations of the PP-EnKF were significantly closer to the reference than the correlations of the classical EnKF. All in all, the performance of the PP-EnKF makes it an attractive alternative to existing EnKF methods, especially for large-scale applications using a small ensemble size.

Outlook: The PP-EnKF can be further tested and applied to more complicated parameter estimation experiments. Regarding testing, it would be beneficial to further test the performance of the filter for various, possibly heterogeneous, pilot point grid configurations. Another interesting research direction could be optimizing the pilot point locations using optimal experimental design strategies. Regarding possible applications for the PP-EnKF, I propose three-dimensional and real-world simulations. On the basis of the results from this work, I predict that the PP-EnKF will be computationally beneficial in large-scale simulations. In particular, the PP-EnKF is expected to be fast, when the parameter estimation is applied to a limited portion of a large model. Often groundwater hydrological models or geothermal reservoir engineering models meet this condition, as large subsurface models are computed, while information is only available at a limited number of locations, for example around wells or boreholes. According to the distribution of information, the pilot point grid can be adjusted to the locations that are known to yield additional information. This makes the PP-EnKF well adapted for such simulations.

Overall conclusions Two general conclusions are now listed that the author draws from the work on this thesis: (1) for scientific software development, (2) for EnKF method development, in particular in groundwater flow estimation. Regarding scientific software, this work supports the conclusions that are most probably already drawn from many other works: Modularity is very important in order to maintain clarity of the code. This is especially important for parameter estimation software, where scientific codes of different disciplines are merged. In the authors view, more time and funding should be invested into refactoring existing scientific software for making it easily accessible. Concretely, this refactoring could include two points:

First, modern scientific software, including for example the usage of version control, should be turned into its own documentation. Secondly, scientific software should contain citations and method explanations comparable to the ones found in scientific publications. Regarding EnKF development, the main conclusion from this work, as stated before, is that comparison of methods should be enhanced. In general, the work on this thesis would have benefited from a set of existing benchmark permeability estimations. Such benchmarks are the only way to ensure true reproducibility and also comparability between several studies. Thus, a suite of benchmarks should be initiated. In the field of parameter estimation, growing computational resources will lead to the possibility of using more refined estimation methods, for example moving from Kalman filtering to particle filtering. However, in the author's view there will always be the need for efficient methods like the EnKF, since the physical simulation models will keep growing alongside the computational resources as well.

Outlook: The future of numerical and inverse groundwater flow modeling will most certainly be shaped by the growth of computational power. On the basis of the method comparisons of this thesis, I propose that part of this growing computational power should be used for robust performance comparisons, for example by carrying out multiple synthetic experiments in these performance comparisons. Secondly, I propose the PP-EnKF as an efficient method for parameter estimation in large-scale groundwater flow models that reduces spurious correlations. Finally, a third proposition following from this work concerns scientific software development. For scientific software development, I propose the usage of the flexible git branches workflow that has been applied to SHEMAT-Suite in this thesis. This workflow can help to develop scientific software that is stable, and at the same time fast, easily manipulable, and readable.

Bibliography

Aanonsen, S. I., Nævdal, G., Oliver, D. S., Reynolds, A. C., and Vallès, B. (2009). The ensemble Kalman filter in reservoir engineering–A review. *SPE Journal*, **14**(03), 393–412. http://dx.doi.org/10.2118/117274-pa

Ahrens, J., Geveci, B., and Law, C. (2005). Paraview: An end-user tool for large data visualization. *The visualization handbook, Elsevier*, **717**(8).

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Anderson, J. L. (2001). An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review*, **129**(12), 2884–2903. http://dx.doi.org/10.1175/1520-0493(2001)129<2884:aeakff>2.0.co;2

Arnold, V. I. (2004). *Lectures on Partial Differential Equations*. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-662-05441-3

Aster, R. C., Borchers, B., and Thurber, C. H. (2013). *Parameter Estimation and Inverse Problems*. Elsevier. http://dx.doi.org/10.1016/C2009-0-61134-X

Baatz, D., Kurtz, W., Hendricks Franssen, H., Vereecken, H., and Kollet, S. (2017). Catchment tomography - An approach for spatial parameter estimation. *Advances in Water Resources*, **107**, 147–159. http://dx.doi.org/10.1016/j.advwatres.2017.06.006

Balay, S., Gropp, W. D., McInnes, L. C., and Smith, B. F. (1997). Efficient management of parallelism in object-oriented numerical software libraries. *Modern Software Tools for Scientific Computing*, page 163–202. http://dx.doi.org/10.1007/978-1-4612-1986-6_8

Batu, V. (2010). *Aquifer Hydraulics: A Comprehensive Guide to Hydrogeologic Data Analysis*. Wiley, New York. https://www.wiley.com/en-us/Aquifer+Hydraulics%3A+A+Comprehensive+Guide+to+Hydrogeologic+Data+Analysis-p-9780471185024

Bear, J. (1975). Dynamics of fluids in porous media. *Soil Science*, **120**(2), 162–163. http://dx.doi.org/10.1097/00010694-197508000-00022

Bear, J., Tsang, C.-F., and Marsily, G. de (1993). *Flow and Contaminant Transport in Fractured Rock*. Elsevier. http://dx.doi.org/10.1016/c2009-0-29127-6

Bear, J. and Verruijt, A. (1987). *Modeling Groundwater Flow and Pollution*. Springer Netherlands.
http://dx.doi.org/10.1007/978-94-009-3379-8

Bocquet, M. and Sakov, P. (2013). An iterative ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, **140**(682), 1521–1535.
http://dx.doi.org/10.1002/qj.2236

Bruckmann, J. and Clauser, C. (2020). Ensemble-based stochastic permeability and flow simulation of a sparsely sampled hard-rock aquifer supported by high performance computing. *Hydrogeology Journal*.
http://dx.doi.org/10.1007/s10040-020-02163-5

Bücker, H. M., Rasch, A., Rath, V., and Wolf, A. (2009). Semi-automatic parallelization of direct and inverse problems for geothermal simulation. In *Proceedings of the 24th ACM Symposium on Applied Computing, Honolulu, Hawaii, USA, March 8–12, 2009*, vol. 2, pages 971–975. ACM Press, New York.
https://doi.org/10.1145/1529282.1529495

Büsing, H., Willkomm, J., Bischof, C. H., and Clauser, C. (2014). Using exact Jacobians in an implicit Newton method for solving multiphase flow in porous media. *International Journal of Computational Science and Engineering*, **9**(5/6), 499.
http://dx.doi.org/10.1504/ijcse.2014.064535

Büsing, H., Niederau, J., and Clauser, C. (2016). Pressure-enthalpy formulation for numerical simulations of supercritical water/steam systems applied to a reservoir in Tuscany, Italy. In *European Geothermal Congress 2016 Strasbourg, France*, pages 1–6.

Büsing, H., Vogt, C., Ebigbo, A., and Klitzsch, N. (2017). Numerical study on CO2 leakage detection using electrical streaming potential data. *Water Resources Research*, **53**(1), 455–469.
http://dx.doi.org/10.1002/2016wr019803

Büsing, H. (2020). Efficient solution techniques for two-phase flow in heterogeneous porous media using exact Jacobians. *Computational Geosciences*. In Review.

Burgers, G., Jan van Leeuwen, P., and Evensen, G. (1998). Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, **126**(6), 1719–1724.
http://dx.doi.org/10.1175/1520-0493(1998)126<1719:asitek>2.0.co;2

Camporese, M., Paniconi, C., Putti, M., and Salandin, P. (2009). Comparison of data assimilation techniques for a coupled model of surface and subsurface flow. *Vadose Zone Journal*, **8**(4), 837.
http://dx.doi.org/10.2136/vzj2009.0018

Cao, Z., Li, L., and Chen, K. (2018). Bridging iterative ensemble smoother and multiple-point geostatistics for better flow and transport modeling. *Journal of Hydrology*, **565**, 411–421.
http://dx.doi.org/10.1016/j.jhydrol.2018.08.023

Capilla, J. E., Jaime Gómez-Hernández, J., and Sahuquillo, A. (1997). Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data 2. Demonstration on a synthetic aquifer. *Journal of Hydrology*, **203**(1-4), 175–188.
http://dx.doi.org/10.1016/s0022-1694(97)00097-8

Carrera, J. and Neuman, S. P. (1986). Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resources Research*, **22**(2), 199–210.
http://dx.doi.org/10.1029/wr022i002p00199

Carslaw, H. S. and Jaeger, J. C. (1959). *Conduction of Heat in Solids*. Oxford: Clarendon Press, 1959, 2nd ed.

Chaudhuri, A., Franssen, H.-J. H., and Sekhar, M. (2018). Iterative filter based estimation of fully 3D heterogeneous fields of permeability and Mualem-van Genuchten parameters. *Advances in Water Resources*, **122**, 340–354.
http://dx.doi.org/10.1016/j.advwatres.2018.10.023

Chen, T., Clauser, C., Marquart, G., Willbrand, K., and Büsing, H. (2016). Modeling anisotropic flow and heat transport by using mimetic finite differences. *Advances in Water Resources*, **94**, 441–456.
http://dx.doi.org/10.1016/j.advwatres.2016.06.006

Chen, T., Clauser, C., Marquart, G., Willbrand, K., and Hiller, T. (2018). Upscaling permeability for three-dimensional fractured porous rocks with the multiple boundary method. *Hydrogeology Journal*, **26**(6), 1903–1916.
http://dx.doi.org/10.1007/s10040-018-1744-z

Chen, Y. and Zhang, D. (2006). Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Advances in Water Resources*, **29**(8), 1107–1122.
http://dx.doi.org/10.1016/j.advwatres.2005.09.007

Chen, Y. and Oliver, D. S. (2010). Cross-covariances and localization for EnKF in multiphase flow data assimilation. *Computational Geosciences*, **14**(4), 579–601.
http://dx.doi.org/10.1007/s10596-009-9174-6

Chen, Y. and Oliver, D. S. (2011). Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences*, **44**(1), 1–26.
http://dx.doi.org/10.1007/s11004-011-9376-z

Chen, Z. (2003). Bayesian filtering: From Kalman filters to particle filters, and beyond. *Statistics*, **182**(1), 1–69.

Chorin, A. J. and Marsden, J. E. (1979). *A Mathematical Introduction to Fluid Mechanics*. Springer US.
http://dx.doi.org/10.1007/978-1-4684-0082-3

Christie, D. E. and Aris, R. (1965). Vectors, tensors, and the basic equations of fluid mechanics. *The American Mathematical Monthly*, **72**(1), 97.
http://dx.doi.org/10.2307/2313035

Clauser, C. (1988). *Untersuchungen zur Trennung der konduktiven und konvektiven Anteile im Wärmetransport in einem Sedimentbecken am Beispiel des Oberrheintalgrabens*. Ph.D. thesis, Technische Universität Berlin, Düsseldorf.

Clauser, C. (1997). Geothermal energy use in Germany—Status and potential. *Geothermics*, **26**(2), 203–220.

Clauser, C. (ed.). (2003). *Numerical Simulation of Reactive Flow in Hot Aquifers: SHEMAT and Processing SHEMAT*. Springer-Verlag Berlin Heidelberg New York.
http://dx.doi.org/10.1007/978-3-642-55684-5

Cosme, E., Verron, J., Brasseur, P., Blum, J., and Auroux, D. (2012). Smoothing problems in a Bayesian framework and their linear Gaussian solutions. *Monthly Weather Review*, **140**(2), 683–695.
http://dx.doi.org/10.1175/mwr-d-10-05025.1

Courant, R. and Hilbert, D. (1924). *Methoden der mathematischen Physik*. Springer, Berlin, Heidelberg.
http://dx.doi.org/10.1007/978-3-642-58039-0

Crank, J. and Nicolson, P. (1947). A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Mathematical Proceedings of the Cambridge Philosophical Society*, **43**(1), 50–67.

Dake, L. P. (1983). *Fundamentals of Reservoir Engineering*. Elsevier.

De Lannoy, G. J. M. and Reichle, R. H. (2016). Assimilation of SMOS brightness temperatures or soil moisture retrievals into a land surface model. *Hydrology and Earth System Sciences*, **20**(12), 4895–4911.
http://dx.doi.org/10.5194/hess-20-4895-2016

Deb, P., Knapp, D., Clauser, C., and Montegrossi, G. (2019). Modeling natural steady-state of super hot geothermal reservoir at Los Humeros, Mexico. In *European Geothermal Congress 2019 Proceedings*, pages 1–6.

Deutsch, C. and Journel, A. (1995). *GSLIB: Geostatistical Software Library and User's Guide*, vol. 37. University Press, Oxford.
http://dx.doi.org/10.2307/1269177

Devegowda, D., Arroyo-Negrete, E., and Datta-Gupta, A. (2010). Flow relevant covariance localization during dynamic data assimilation using EnKF. *Advances in Water Resources*, **33**(2), 129–145.
http://dx.doi.org/10.1016/j.advwatres.2009.10.001

Diersch, H.-J. (2014). *FEFLOW: Finite Element Modeling of Flow, Mass and Heat Transport in Porous and Fractured Media*. Springer-Verlag Berlin Heidelberg.
http://dx.doi.org/10.1007/978-3-642-38739-5

Dongarra, J. J., Du Croz, J., Hammarling, S., and Duff, I. S. (1990). A Set of Level 3 Basic Linear Algebra Subprograms. *ACM Trans. Math. Softw.*, **16**(1), 1–17.
https://doi.org/10.1145/77626.79170

Dongarra, J. J., Du Croz, J., Hammarling, S., and Hanson, R. J. (1988). An Extended Set of FORTRAN Basic Linear Algebra Subprograms. *ACM Trans. Math. Softw.*, **14**(1), 1–17.
https://doi.org/10.1145/42288.42291

Ebigbo, A., Niederau, J., Marquart, G., Dini, I., Thorwart, M., Rabbel, W., Pechnig, R., Bertani, R., and Clauser, C. (2016). Influence of depth, temperature, and structure of a crustal heat source on the geothermal reservoirs of Tuscany: Numerical modelling and sensitivity study. *Geothermal Energy*, **4**(1).
http://dx.doi.org/10.1186/s40517-016-0047-7

El Gharamti, M., Hoteit, I., and Valstar, J. (2013). Dual states estimation of a subsurface flow-transport coupled model using ensemble Kalman filtering. *Advances in Water Resources*, **60**, 75–88.
http://dx.doi.org/10.1016/j.advwatres.2013.07.011

El Gharamti, M., Valstar, J., and Hoteit, I. (2014). An adaptive hybrid EnKF-OI scheme for efficient state-parameter estimation of reactive contaminant transport models. *Advances in Water Resources*, **71**, 1–15.
http://dx.doi.org/10.1016/j.advwatres.2014.05.001

El Gharamti, M. and Hoteit, I. (2014). Complex step-based low-rank extended Kalman filtering for state-parameter estimation in subsurface transport models. *Journal of Hydrology*, **509**, 588–600.
http://dx.doi.org/10.1016/j.jhydrol.2013.12.004

El Gharamti, M., Ait-El-Fquih, B., and Hoteit, I. (2015). An iterative ensemble Kalman filter with one-step-ahead smoothing for state-parameters estimation of contaminant transport models. *Journal of Hydrology*, **527**, 442–457.
http://dx.doi.org/10.1016/j.jhydrol.2015.05.004

Emerick, A. and Reynolds, A. (2011). Combining sensitivities and prior information for covariance localization in the ensemble Kalman filter for petroleum reservoir applications. *Computational Geosciences*, **15**(2), 251–269.
http://dx.doi.org/10.1007/s10596-010-9198-y

Emerick, A. A. and Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, **55**, 3–15.
http://dx.doi.org/10.1016/j.cageo.2012.03.011

Evans, L. (2010). *Partial Differential Equations*. American Mathematical Society. http://dx.doi.org/10.1090/gsm/019

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99**(C5), 10143. http://dx.doi.org/10.1029/94jc00572

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, **53**(4), 343–367. http://dx.doi.org/10.1007/s10236-003-0036-9

Evensen, G. (2009). The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems Magazine*, **29**(3), 83–104. http://dx.doi.org/10.1109/mcs.2009.932223

Evensen, G. and van Leeuwen, P. J. (2000). An ensemble Kalman smoother for nonlinear dynamics. *Monthly Weather Review*, **128**(6), 1852–1867. http://dx.doi.org/10.1175/1520-0493(2000)128<1852:aeksfn>2.0.co;2

Finsterle, S., Sonnenthal, E. L., and Spycher, N. (2014). Advances in subsurface modeling using the TOUGH suite of simulators. *Computers & Geosciences*, **65**, 2–12. http://dx.doi.org/10.1016/j.cageo.2013.06.009

Flemisch, B., Darcis, M., Erbertseder, K., Faigle, B., Lauser, A., Mosthaf, K., Müthing, S., Nuske, P., Tatomir, A., Wolff, M., and et al. (2011). DuMu^{x}: DUNE for multi-{phase, component, scale, physics, ...} flow and transport in porous media. *Advances in Water Resources*, **34**(9), 1102–1112. http://dx.doi.org/10.1016/j.advwatres.2011.03.007

Gaspari, G. and Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, **125**(554), 723–757. http://dx.doi.org/10.1002/qj.49712555417

Gelb, A. (1974). *Applied Optimal Estimation*. The MIT press.

Gerke, K. M., Vasilyev, R. V., Khirevich, S., Collins, D., Karsanina, M. V., Sizonenko, T. O., Korost, D. V., Lamontagne, S., and Mallants, D. (2018). Finite-difference method Stokes solver (FDMSS) for 3D pore geometries: Software development, validation and case studies. *Computers & Geosciences*, **114**, 41–58. http://dx.doi.org/10.1016/j.cageo.2018.01.005

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford university press.

Grant, M. (2013). *Geothermal Reservoir Engineering*. Elsevier.

Gómez-Hernández, J., Sahuquillo, A., and Capilla, J. (1997). Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data—I. Theory. *Journal of*

Hydrology, **203**(1-4), 162–174.
http://dx.doi.org/10.1016/s0022-1694(97)00098-x

Hamill, T. M. and Snyder, C. (2000). A hybrid ensemble Kalman filter–3D variational analysis scheme. *Monthly Weather Review*, **128**(8), 2905–2919.
http://dx.doi.org/10.1175/1520-0493(2000)128<2905:ahekfv>2.0.co;2

Hamill, T. M., Whitaker, J. S., and Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, **129**(11), 2776–2790.
http://dx.doi.org/10.1175/1520-0493(2001)129<2776:ddfobe>2.0.co;2

Hascoët, L. and Pascual, V. (2013). The Tapenade automatic differentiation tool: Principles, model, and specification. *ACM Transactions on Mathematical Software*, **39**(3), 20:1–20:43.
http://dx.doi.org/10.1145/2450153.2450158

Heidari, L., Gervais, V., Ravalec, M. L., and Wackernagel, H. (2013). History matching of petroleum reservoir models by the ensemble Kalman filter and parameterization methods. *Computers & Geosciences*, **55**, 84–95.
http://dx.doi.org/10.1016/j.cageo.2012.06.006

Hendricks Franssen, H.-J., Gómez-Hernández, J., and Sahuquillo, A. (2003). Coupled inverse modelling of groundwater flow and mass transport and the worth of concentration data. *Journal of Hydrology*, **281**(4), 281–295.
http://dx.doi.org/10.1016/s0022-1694(03)00191-4

Hendricks Franssen, H. J. and Kinzelbach, W. (2008). Real-time groundwater flow modeling with the ensemble Kalman filter: Joint estimation of states and parameters and the filter inbreeding problem. *Water Resources Research*, **44**(9).
http://dx.doi.org/10.1029/2007wr006505

Hendricks Franssen, H. and Kinzelbach, W. (2009). Ensemble Kalman filtering versus sequential self-calibration for inverse modelling of dynamic groundwater flow systems. *Journal of Hydrology*, **365**(3-4), 261–274.
http://dx.doi.org/10.1016/j.jhydrol.2008.11.033

Houtekamer, P. L. and Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, **126**(3), 796–811.
http://dx.doi.org/10.1175/1520-0493(1998)126<0796:dauaek>2.0.co;2

Houtekamer, P. L. and Zhang, F. (2016). Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **144**(12), 4489–4532.
http://dx.doi.org/10.1175/mwr-d-15-0440.1

Hruška, M., Clauser, C., and De Doncker, R. W. (2018). The effect of drying around power cables on the vadose zone temperature. *Vadose Zone Journal*, **17**(1), 180105.
http://dx.doi.org/10.2136/vzj2018.05.0105

Hruška, M., Clauser, C., and De Doncker, R. W. (2019). Influence of dry ambient conditions on performance of underground medium-voltage DC cables. *Applied Thermal Engineering*, **149**, 1419–1426.
http://dx.doi.org/10.1016/j.applthermaleng.2018.11.076

Hughes, J. D., Langevin, C. D., and Banta, E. R. (2017). Documentation for the MODFLOW 6 framework. *Techniques and Methods*.
http://dx.doi.org/10.3133/tm6a57

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, **9**(3), 90–95.

Huyakorn, P. S. (2012). *Computational Methods in Subsurface Flow*. Academic Press.

Hwang, L., Fish, A., Soito, L., Smith, M., and Kellogg, L. H. (2017). Software and the scientist: Coding and citation practices in geodynamics. *Earth and Space Science*, **4**(11), 670–680.
http://dx.doi.org/10.1002/2016ea000225

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. Courier Corporation.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic press.

Kaipio, J. and Somersalo, E. (2006). *Statistical and Computational Inverse Problems*, vol. 160. Springer Science & Business Media.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **82**(1), 35.
http://dx.doi.org/10.1115/1.3662552

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering*, **83**(1), 95.
http://dx.doi.org/10.1115/1.3658902

Kalnay, E. (2002). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press.
http://dx.doi.org/10.1017/cbo9780511802270

Kang, B., Yang, H., Lee, K., and Choe, J. (2017). Ensemble Kalman filter with principal component analysis assisted sampling for channelized reservoir characterization. *Journal of Energy Resources Technology*, **139**(3).
http://dx.doi.org/10.1115/1.4035747

Kinzelbach, W. and Rausch, R. (1995). *Grundwassermodellierung*. Schweizerbart'sche Verlagsbuchhandlung.

Kitanidis, P. K. and Vomvoris, E. G. (1983). A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations. *Water Resources Research*, **19**(3), 677–690.
http://dx.doi.org/10.1029/wr019i003p00677

Kolditz, O., Bauer, S., Bilke, L., Böttcher, N., Delfs, J. O., Fischer, T., Görke, U. J., Kalbacher, T., Kosakowski, G., McDermott, C. I., and et al. (2012). OpenGeoSys: An open-source initiative for numerical simulation of thermo-hydro-mechanical/chemical (THM/C) processes in porous media. *Environmental Earth Sciences*, **67**(2), 589–599.
http://dx.doi.org/10.1007/s12665-012-1546-x

Kurtz, W., Hendricks Franssen, H.-J., Kaiser, H.-P., and Vereecken, H. (2014). Joint assimilation of piezometric heads and groundwater temperatures for improved modeling of river-aquifer interactions. *Water Resources Research*, **50**(2), 1665–1688.
http://dx.doi.org/10.1002/2013wr014823

Kürten, S., Mottaghy, D., and Ziegler, M. (2014). A new model for the description of the heat transfer for plane thermo-active geotechnical systems based on thermal resistances. *Acta Geotechnica*, **10**(2), 219–229.
http://dx.doi.org/10.1007/s11440-014-0311-6

Kürten, S., Mottaghy, D., and Ziegler, M. (2015). Design of plane energy geostructures based on laboratory tests and numerical modelling. *Energy and Buildings*, **107**, 434–444.
http://dx.doi.org/10.1016/j.enbuild.2015.08.039

Lawson, C. L., Hanson, R. J., Kincaid, D. R., and Krogh, F. T. (1979). Basic Linear Algebra Subprograms for Fortran Usage. *ACM Trans. Math. Softw.*, **5**(3), 308–323.
https://doi.org/10.1145/355841.355847

Li, L., Srinivasan, S., Zhou, H., and Gómez-Hernández, J. (2013). A pilot point guided pattern matching approach to integrate dynamic data into geological modeling. *Advances in Water Resources*, **62**, 125–138.
http://dx.doi.org/10.1016/j.advwatres.2013.10.008

Li, L., Zhou, H., Hendricks Franssen, H. J., and Gómez-Hernández, J. J. (2012). Groundwater flow inverse modeling in non-multiGaussian media: Performance assessment of the normal-score ensemble Kalman filter. *Hydrology and Earth System Sciences*, **16**(2), 573–590.
http://dx.doi.org/10.5194/hess-16-573-2012

Li, S., Reuning, L., Marquart, G., Wang, Y., and Zhao, P. (2017). Numerical model of halite precipitation in porous sedimentary rocks adjacent to salt diapirs. *Journal of Geophysics and Engineering*, **14**(5), 1160–1166.
http://dx.doi.org/10.1088/1742-2140/aa73f9

Liu, B., El Gharamti, M., and Hoteit, I. (2016). Assessing clustering strategies for Gaussian mixture filtering a subsurface contaminant model. *Journal of Hydrology*, **535**, 1–21.
http://dx.doi.org/10.1016/j.jhydrol.2016.01.048

Lorentzen, R., Nævdal, G., and Lage, A. (2003). Tuning of parameters in a two-phase flow model using an ensemble Kalman filter. *International Journal of Multiphase Flow*, **29**(8), 1283–1309. http://dx.doi.org/10.1016/s0301-9322(03)00088-0

Loève, M. (1977). *Probability Theory I.* Springer Verlag, New York.

Lynch, D. R. (2005). *Numerical Partial Differential Equations for Environmental Scientists and Engineers: A First Practical Course.* Springer Science & Business Media.

Marsily, G. de (1986). *Quantitative Hydrology.* Academic Press, San Diego, Calif.

Michalski, A. and Klitzsch, N. (2018). Temperature sensor module for groundwater flow detection around borehole heat exchangers. *Geothermal Energy*, **6**(1), 15. http://dx.doi.org/10.1186/s40517-018-0101-8

Michalski, A. and Klitzsch, N. (2019). First field application of temperature sensor modules for groundwater flow detection near borehole heat exchanger. *Geothermal Energy*, **7**(1). http://dx.doi.org/10.1186/s40517-019-0152-5

Moradkhani, H., Sorooshian, S., Gupta, H. V., and Houser, P. R. (2005). Dual state–parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources*, **28**(2), 135–147. http://dx.doi.org/10.1016/j.advwatres.2004.09.002

Mottaghy, D. and Dijkshoorn, L. (2012). Implementing an effective finite difference formulation for borehole heat exchangers into a heat and mass transport code. *Renewable Energy*, **45**, 59–71. http://dx.doi.org/10.1016/j.renene.2012.02.013

Mottaghy, D. and Rath, V. (2006). Latent heat effects in subsurface heat transport modelling and their impact on palaeotemperature reconstructions. *Geophysical Journal International*, **164**(1), 236–245. http://dx.doi.org/10.1111/j.1365-246x.2005.02843.x

Naevdal, G., Johnsen, L. M., Aanonsen, S. I., and Vefring, E. H. (2005). Reservoir monitoring and continuous model updating using ensemble Kalman filter. *SPE Journal*, **10**(01), 66–74. http://dx.doi.org/10.2118/84372-pa

Nerger, L. and Hiller, W. (2013). Software for ensemble-based data assimilation systems - Implementation strategies and scalability. *Computers & Geosciences*, **55**, 110–118. http://dx.doi.org/10.1016/j.cageo.2012.03.026

Niederau, J., Ebigbo, A., Marquart, G., Arnold, J., and Clauser, C. (2017). On the impact of spatially heterogenous permeability on free convection in the Perth Basin, Australia. *Geothermics*, **66**, 119–133. http://dx.doi.org/10.1016/j.geothermics.2016.11.011

Nowak, W. (2009). Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator. *Water Resources Research*, **45**(4).
http://dx.doi.org/10.1029/2008wr007328

Pagani, S., Manzoni, A., and Quarteroni, A. (2017). Efficient state/parameter estimation in nonlinear unsteady PDEs by a reduced basis ensemble Kalman filter. *SIAM/ASA Journal on Uncertainty Quantification*, **5**(1), 890–921.
http://dx.doi.org/10.1137/16m1078598

Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables, and Stochastic Processes*. Tata McGraw-Hill Education.

Raanes, P. N., Bocquet, M., and Carrassi, A. (2019). Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. *Quarterly Journal of the Royal Meteorological Society*, **145**(718), 53–75.
http://dx.doi.org/10.1002/qj.3386

Radosavljevic, B., Elger, K., Bertelmann, R., Haberland, C., Hemmleb, S., Muñoz, G., Quinteros, J., and Strollo, A. (2019). Report on the survey of digital data management practices at the GFZ German Research Centre for Geosciences. Tech. rep., Scientific Technical Report STR19/02.

RamaRao, B. S., LaVenue,A. M., Marsily, G. de, and Marietta, M. G. (1995). Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: 1. Theory and computational experiments. *Water Resources Research*, **31**(3), 475–493.
http://dx.doi.org/10.1029/94wr02258

Rasch, A. and Bücker, H. M. (2010). EFCOSS: An interactive environment facilitating optimal experimental design. *ACM Transactions on Mathematical Software*, **37**(2), 13:1–13:37.
https://doi.org/10.1145/1731022.1731023

Rath, V., Wolf, A., and Bücker, H. M. (2006). Joint three-dimensional inversion of coupled groundwater flow and heat transfer based on automatic differentiation: sensitivity calculation, verification, and synthetic examples. *Geophysical Journal International*, **167**(1), 453–466.
http://dx.doi.org/10.1111/j.1365-246x.2006.03074.x

Reichle, R. H., McLaughlin, D. B., and Entekhabi, D. (2002). Hydrologic data assimilation with the ensemble Kalman filter. *Monthly Weather Review*, **130**(1), 103–114.
http://dx.doi.org/10.1175/1520-0493(2002)130<0103:hdawte>2.0.co;2

Rostami, M. A. and Bücker, H. M. (2014). Preservation of non-uniform memory architecture characteristics when going from a nested OpenMP to a hybrid MPI/OpenMP approach. In Obaidat, M. S., Kacprzyk, J., and Ören, T. (eds.) *SIMULTECH 2014, Proceedings of the 4th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, Vienna, Austria, August 28–30, 2014*, pages 286–291. SciTePress.
https://doi.org/10.5220/0005110902860291

Rostami, M. A., Bücker, H. M., Vogt, C., Seidler, R., Neuhäuser, D., and Rath, V. (2014). A distributed-memory parallelization of a shared-memory parallel ensemble Kalman filter. In Winkler, F., Negru, V., Ida, T., Jebelean, T., Petcu, D., Watt, S., and Zaharie, D. (eds.) *Proceedings of the 16th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2014), September 22–25, 2014, Timisoara, Romania*, pages 455–462. IEEE Computer Society, Los Alamitos, CA, USA.
https://doi.org/10.1109/SYNASC.2014.67

Sakov, P., Haussaire, J.-M., and Bocquet, M. (2018). An iterative ensemble Kalman filter in the presence of additive model error. *Quarterly Journal of the Royal Meteorological Society*, **144**(713), 1297–1309.
http://dx.doi.org/10.1002/qj.3213

Sakov, P., Oliver, D. S., and Bertino, L. (2012). An iterative EnKF for strongly nonlinear systems. *Monthly Weather Review*, **140**(6), 1988–2004.
http://dx.doi.org/10.1175/mwr-d-11-00176.1

Schöniger, A., Nowak, W., and Hendricks Franssen, H.-J. (2012). Parameter estimation by ensemble Kalman filters with transformed data: Approach and application to hydraulic tomography. *Water Resources Research*, **48**(4).
http://dx.doi.org/10.1029/2011wr010462

Schroeder, W., Martin, K., and Lorensen, B. (2006). *The Visualization Toolkit*. Kitware.
https://vtk.org/documentation/#textbook

Segal, J. and Morris, C. (2008). Developing scientific software. *IEEE Software*, **25**(4), 18–20.
http://dx.doi.org/10.1109/ms.2008.85

Segol, G. (1994). *Classic Groundwater Simulations: Proving and Improving Numerical Models*. Prentice Hall.

Seidler, R., Padalkina, K., Bücker, H. M., Ebigbo, A., Herty, M., Marquart, G., and Niederau, J. (2016). Optimal experimental design for reservoir property estimates in geothermal exploration. *Computational Geosciences*, **20**(2), 375–383.
http://dx.doi.org/10.1007/s10596-016-9565-4

Shi, Y., Davis, K. J., Zhang, F., Duffy, C. J., and Yu, X. (2014). Parameter estimation of a physically based land surface hydrologic model using the ensemble Kalman filter: A synthetic experiment. *Water Resources Research*, **50**(1), 706–724.
http://dx.doi.org/10.1002/2013wr014070

Sivia, D. and Skilling, J. (2006). *Data Analysis: A Bayesian Tutorial*. Oxford University Press.

Sun, A. Y., Morris, A., and Mohanty, S. (2009a). Comparison of deterministic ensemble Kalman filters for assimilating hydrogeological data. *Advances in Water Resources*, **32**(2), 280–292.
http://dx.doi.org/10.1016/j.advwatres.2008.11.006

Sun, A. Y., Morris, A. P., and Mohanty, S. (2009b). Sequential updating of multimodal hydrogeologic parameter fields using localization and clustering techniques. *Water Resources Research*, **45**(7).
http://dx.doi.org/10.1029/2008wr007443

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial & Applied Mathematics (SIAM).
http://dx.doi.org/10.1137/1.9780898717921

Tavakoli, R., Yoon, H., Delshad, M., ElSheikh, A. H., Wheeler, M. F., and Arnold, B. W. (2013). Comparison of ensemble filtering algorithms and null-space Monte Carlo for parameter estimation and uncertainty quantification using CO2 sequestration data. *Water Resources Research*, **49**(12), 8108–8127.
http://dx.doi.org/10.1002/2013wr013959

The HDF Group (1997-2022). Hierarchical Data Format, version 5.
https://www.hdfgroup.org/HDF5/

Theis, C. V. (1935). The relation between the lowering of the piezometric surface and the rate and duration of discharge of a well using ground-water storage. *Transactions, American Geophysical Union*, **16**(2), 519.
http://dx.doi.org/10.1029/tr016i002p00519

Thomas, A. T., Reiche, S., Riedel, M., and Clauser, C. (2019). The fate of submarine fresh groundwater reservoirs at the New Jersey shelf, USA. *Hydrogeology Journal*, **27**(7), 2673–2694.
http://dx.doi.org/10.1007/s10040-019-01997-y

Trescott, P. and Larson, S. (1977). Solution of three-dimensional groundwater flow equations using the strongly implicit procedure. *Journal of Hydrology*, **35**(1-2), 49–60.
http://dx.doi.org/10.1016/0022-1694(77)90076-2

Trescott, P. C. and Larson, S. (1976). Documentation of finite-difference model for simulation of three-dimensional ground-water flow. *Open-File Report*.
http://dx.doi.org/10.3133/ofr76591

van der Vorst, H. A. (1992). Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, **13**(2), 631–644.
http://dx.doi.org/10.1137/0913035

Vogt, C., Mottaghy, D., Wolf, A., Rath, V., Pechnig, R., and Clauser, C. (2010). Reducing temperature uncertainties by stochastic geothermal reservoir modelling. *Geophysical Journal International*, **181**(1), 321–333.
http://dx.doi.org/10.1111/j.1365-246x.2009.04498.x

Vogt, C., Marquart, G., Kosack, C., Wolf, A., and Clauser, C. (2012). Estimating the permeability distribution and its uncertainty at the EGS demonstration reservoir Soultz-sous-Forêts

using the ensemble Kalman filter. *Water Resources Research*, **48**(8).
http://dx.doi.org/10.1029/2011wr011673

Vogt, C., Klitzsch, N., and Rath, V. (2014). On self-potential data for estimating permeability in enhanced geothermal systems. *Geothermics*, **51**, 201–213.
http://dx.doi.org/10.1016/j.geothermics.2014.01.008

Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., and Verstraten, J. M. (2005). Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resources Research*, **41**(1).
http://dx.doi.org/10.1029/2004wr003059

Wan, E. A. and Nelson, A. T. (2001). Dual extended Kalman filter methods. *Kalman Filtering and Neural Networks*, page 123–173.
http://dx.doi.org/10.1002/0471221546.ch5

White, M. D., Oostrom, M., Rockhold, M. L., and Rosing, M. (2008). Scalable modeling of carbon tetrachloride migration at the Hanford site using the STOMP simulator. *Vadose Zone Journal*, **7**(2), 654–666.
http://dx.doi.org/10.2136/vzj2007.0070

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., and et al. (2014). Best practices for scientific computing. *PLoS Biology*, **12**(1), e1001745.
http://dx.doi.org/10.1371/journal.pbio.1001745

Wolf, A. (2011). *Ein Softwarekonzept zur hierarchischen Parallelisierung von stochastischen und deterministischen Inversionsproblemen auf modernen ccNUMA-Plattformen unter Nutzung automatischer Programmtransformation.* Ph.D. thesis, RWTH Aachen University.

Wolf, A., Rath, V., and Bücker, H. M. (2008). Parallelisation of a geothermal simulation package: A case study on four multicore architectures. In Bischof, C., Bücker, M., Gibbon, P., Joubert, G., Lippert, T., Mohr, B., and Peters, F. (eds.) *Parallel Computing: Architectures, Algorithms and Applications*, vol. 15 of *Advances in Parallel Computing*, pages 451–458. IOS Press, Amsterdam, The Netherlands. ISBN 978-1-58603-796-3. Co-published with NIC-Directors as Volume 38 of NIC Series, 2007.
http://hdl.handle.net/2128/6498

Wu, C.-C. and Margulis, S. A. (2011). Feasibility of real-time soil state and flux characterization for wastewater reuse using an embedded sensor network data assimilation approach. *Journal of Hydrology*, **399**(3-4), 313–325.
http://dx.doi.org/10.1016/j.jhydrol.2011.01.011

Xiao, D., Du, J., Fang, F., Pain, C., and Li, J. (2018). Parameterised non-intrusive reduced order methods for ensemble Kalman filter data assimilation. *Computers & Fluids*, **177**, 69–77.
http://dx.doi.org/10.1016/j.compfluid.2018.10.006

Zheng, C. and Bennett, G. D. (2002). *Applied Contaminant Transport Modeling*, vol. 2. Wiley-Interscience New York.

Zhou, H., Gómez-Hernández, J. J., Hendricks Franssen, H.-J., and Li, L. (2011). An approach to handling non-gaussianity of parameters and state variables in ensemble Kalman filtering. *Advances in Water Resources*, **34**(7), 844–864. http://dx.doi.org/10.1016/j.advwatres.2011.04.014

Zovi, F., Camporese, M., Hendricks Franssen, H.-J., Huisman, J. A., and Salandin, P. (2017). Identification of high-permeability subsurface structures with multiple point geostatistics and normal score ensemble Kalman filter. *Journal of Hydrology*, **548**, 208–224. http://dx.doi.org/10.1016/j.jhydrol.2017.02.056

A

SHEMAT-Suite input files

In this appendix, the input files for SHEMAT-Suite referred to in the main text are collected.

Forward mode

Input File A.1: Input file of the THEIS model.

```
! # =====>>>>>   RUN INFO                              <<<<<===== # !
# title
theis
# linfo
2 3 2 3
# runmode
0
# active head
# PROPS = const
# USER = none
! # =====>>>>>   GRID                                 <<<<<===== # !
# grid
69 69 1
# delx
3*160.0d0 1*80.0d0 1*40.0d0 3*20.0d0 3*10.0d0
5*5.0d0 5*2.0d0 11*1.0d0 2*1.0d0 1.0d0 2*1.0d0
11*1.0d0 5*2.0d0 5*5.0d0 3*10.0d0 3*20.0d0
1*40.0d0 1*80.0d0 3*160.0d0
# dely
3*160.0d0 1*80.0d0 1*40.0d0 3*20.0d0 3*10.0d0
5*5.0d0 5*2.0d0 11*1.0d0 2*1.0d0 1.0d0 2*1.0d0
11*1.0d0 5*2.0d0 5*5.0d0 3*10.0d0 3*20.0d0
1*40.0d0 1*80.0d0 3*160.0d0
# delz
20.d0
! # =====>>>>>   TIME STEP CONTROL                   <<<<<===== # !
# timestep control
1
1.0 1.0 1.0 0.0
# tunit
86400
# time periods, records=1
0.0   3.5     100 lin
! # =====>>>>>   SOLVER CONTROL                      <<<<<===== # !
# nlsolve
50 0
# lsolvef (linear solver control)
1.d-10 64 500
```

```
# nliterf (nonlinear iteration control)
1.0d-10 1.0d0
! # ====>>>>>   BOUNDARY AND INITIAL CONDITIONS    <<<<<===== # !
# head bcn, records=1, direction=0
35 35 1 -0.001d0 0
# head bcd, simple=right, value=init
# head bcd, simple=back, value=init
# head init
4761*20.0d0
# temp init
4761*20.0d0
! # ====>>>>>   PROPERTIES                          <<<<<===== # !
# fluid props
1000.0d0 5.0D-8 4218.0d0 0.65d0 1.0D-3 0.02d0
# units
0.1d0  1.d0  1.d0 1d-12 1.d-8  1.d0  1.d0  3.35d0  0.0d-06
2300000.d0  10d0  0.d0  0.d0
! # ====>>>>>   GEOMETRY                           <<<<<===== # !
# uindex
4761*1
! # ====>>>>>   OUTPUT                             <<<<<===== # !
# file output: vtk hdf
```

Inverse mode

Input File A.2: First part of the input file for the inverse simulation showing specifications for the flow and heat transport model. `runmode 2` defines the inverse simulation. Under `# unit`, the line breaks are only for visualization, in the SHEMAT-Suite input file, one should enter one line per unit. The (...) at the end indicates that the input file is cut for better visualization.

```
!==========>>>>> SHEMAT-Suite Input file <<<<<==========!
# title
demo_inversion
# linfo
1 1 1 1
# runmode
2
# USER=none
# PROPS=bas
# active temp head

!==========>>>>> MESH in meters
# grid
150 1 75
# delx
150*2
# dely
1*2
# delz
75*2

!==========>>>>> I/O
# file output vtk hdf5
# borehole logs, records=3
50  1  L1
75  1  L2
100 1  L3
```

```
!==========>>>>> NONLINEAR SOLVER
# nlsolve
100 0

!==========>>>>> FLOW
# lsolvef (linear solver control)
1.d-8 64 300
# nliterf (nonlinear iteration control)
1.0d-6 1.

!==========>>>>> TEMPERATURE
# lsolvet (linear solver control)
1.d-6 64 300
# nlitert (nonlinear iteration control)
1.0d-4 1.

!==========>>>>> STRUCTURE
# units
0.1e-9  1.d0  1.d0 1.00e-20 1.e-10  1.d0  1.d0 2.00  0.0e-06
2300000. 10d0  0.d0  0.d0 2.0 1.0d6 0.05 0.2  !Unit1: Luft
0.20  1.d0  1.d0 1.00e-16 1.e-10  1.d0  1.d0  2.50  0.0e-06
2300000.  10d0  0.d0  0.d0 2.0 1.0d6 0.05 0.2 !Unit2
0.20  1.d0  1.d0 1.00e-16 1.e-10  1.d0  1.d0  2.00  0.0e-06
2300000.  10d0  0.d0  0.d0 2.0 1.0d6 0.05 0.2 !Unit3
0.20  1.d0  1.d0 1.00e-16 1.e-10  1.d0  1.d0  2.00  0.0e-06
2300000.  10d0  0.d0  0.d0 2.0 1.0d6 0.05 0.2 !Unit4
0.25  1.d0  1.d0 1.00e-14 1.e-10  1.d0  1.d0  2.00  0.0e-06
2300000.  10d0  0.d0  0.d0 2.0 1.0d6 0.05 0.2 !Unit5
0.33  1.d0  1.d0 1.00e-16 1.e-10  1.d0  1.d0  2.00  0.0e-06
2300000.  10d0  0.d0  0.d0 2.0 1.0d6 0.05 0.2 !Unit6
0.20  1.d0  1.d0 1.00e-16 1.e-10  1.d0  1.d0  2.00  0.0e-06
2300000.  10d0  0.d0  0.d0 2.0 1.0d6 0.05 0.2 !Unit7
0.10  1.d0  1.d0 1.00e-18 1.e-10  1.d0  1.d0  2.00  0.0e-06
2300000.  10d0  0.d0  0.d0 2.0 1.0d6 0.05 0.2 !Unit8
0.10  1.d0  1.d0 1.00e-18 1.e-10  1.d0  1.d0  2.00  0.0e-06
2300000.  10d0  0.d0  0.d0 2.0 1.0d6 0.05 0.2 !Unit9

# uindex
895*9 5*8 125*9 25*8 105*9 45*8 86*9 64*8 66*9 84*8 46*9 104*8
27*9 123*8 7*9 741*8 2*7 130*8 20*7 111*8 39*7 93*8 57*7 74*8
76*7 55*8 95*7 37*8 97*7 16*6 18*8 98*7 34*6 97*7 53*6 79*7 71*6
60*7 90*6 42*7 108*6 23*7 127*6 5*7 2236*6 9*4 116*6 34*4 90*6
60*4 65*6 85*4 40*6 110*4 15*6 727*4 8*5 122*4 28*5 103*4 47*5
84*4 66*5 64*4 86*5 45*4 105*5 25*4 125*5 722*3 28*1 115*3 35*1
95*3 55*1 92*2 58*1 89*2 61*1 82*2 68*1 68*2 82*1 63*2 87*1 58*2
92*1 52*2 98*1 33*2 117*1 22*2 128*1


!==========>>>>> INITIAL VALUES
# temp init
11250*11.0
# head init
11250*150.0

!==========>>>>> BOUNDARY CONDITIONS
! basal heat flow condition
# temp bcn, simple=base
150*0.06
```

```
! top boundary follows topography
# temp bcd, records=150, error=ignore
1     1        75       11       0
2     1        75       11       0
3     1        75       11       0
4     1        75       11       0
5     1        75       11       0
6     1        75       11       0
7     1        75       11       0
(...)
```

Input File A.3: Continuation of the input file for the inverse simulation showing specifications for the inverse model. The (...) at the end indicates that the file is cut for better visualization. Here, more observation data records have been removed.

```
# inverse
1
5 0.0001 1.0
0 0
1

# errors
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

# enable unit
0 0 0 1 1 1 1 1 1

# enable property
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0

# optimize property, records=6
4 1 lz
5 1 lz
6 1 lz
7 1 lz
8 1 lz
9 1 lz

# data, records=60
16.02 0.1     75      1       6       2       1
16.06 0.1     75      1       7       2       1
16.04 0.1     75      1       8       2       1
16    0.1     75      1       9       2       1
(...)
```

Stochastic mode

Input File A.4: General SHEMAT-Suite input file for the synthetic parameter estimation TRACER.

```
!=========================================================!
!==============    SHEMAT-Suite Inputfile    ============!
!=========================================================!
# Title
TRACER
# linfo
4*0
# runmode
3
# PROPS=const
# USER=none

!=========>>>>>
!=========>>>>>      SIMULATION
!=========>>>>>
# simulate
1000
# enkf iteration
1
# enkf postcompute none
!none mean samples

!==========>>>>>
!==========>>>>>    ROCK MATRIX
!==========>>>>>
# parameter group, records=1
1 parfile=sgsim_k_wavereal.par
1 4 log
# standard deviation
0.0d0 0.0d0 0.0d0 2.302585093d+000 0.0d0 0.0d0 0.0d0 0.0d0 0.0d0
0.0d0 0.0d0 0.0d0 0.0d0 0.0d0 0.0d0 0.0d0 0.0d0
# units
0.1d0  1.d0  1.d0  3.1632d-10 1.d-9  1.d0 1.d0 3.35d0  0.0d-06
2300000.d0  5.0d0  0.d0  0.d0 0.0d0 0.0d0 0.0d0 0.0d0
# split units, records=1
1
# uindex
961*1
```

Input File A.5: SHEMAT-Suite EnKF input file for the synthetic parameter estimation TRACER.

```
!=================================================!
!=-====      EnKF Inputfile          ============!
!=================================================!

!-----------------------------------------!
! Head  ! Temp ! Conc ! Perm ! Tcon ! Poro !
!-----------------------------------------!
!   1   !   2  !   3  !   4  !   5  !   6  !
!-----------------------------------------!

# nrobs_int
10
```

```
# state vector activity
1 0 1 1 0 0

# observation variances
0.0d+0  0.0d+0  50.0d-6   0.0d+0  0.0d0  0.0d0

!=========>>>>>        INPUT

# checktrue
T

# true file names
TrueTRACER.plt
TrueTRACER_chem.plt

# ex_unit
0

# observation file names
observations_TRACER.dat
```

List of symbols

a

EnKF method in a synthetic experiment used for a method comparison 55

$a \in \mathbb{R}^+$

input parameter of the correlation function ρ in the local EnKF 49

$a \in \mathbb{R}^+$

correlation length used for the simulation of permeability fields 52, 53

b

EnKF method in a synthetic experiments used for a method comparison 55

$c\ [ML^{-3}]$

tracer concentration in the species transport equation 11

$c_f\ [L^2T^{-2}\Theta^{-1}]$

fluid specific heat capacity in the heat transport equation 11, 12

$c_m\ [L^2T^{-2}\Theta^{-1}]$

matrix specific heat capacity in the heat transport equation 11, 12

$c_{obs}\ [N/L^3]$

random variable for an observed tracer concentration 84

Cov

sample covariance function 84

$\mathbf{D}\ [L^2T^{-1}]$

hydrodynamic dispersion tensor in the species transport equation 11

$d \in \mathbb{R}^+\ [L]$

distance between two locations in the computation of the correlation matrix ρ of the local EnKF 48, 49

$d \in \mathbb{R}^+$

Euclidean distance between two grid cells used for permeability estimation 52

$\mathbf{d} \in \mathbb{R}^{n_m}$

mean vector of observations in the Kalman filter 18, 75, 77

$d^{a<b,\,n_e} \in [0,1]$

relative difference of RMSE means $\bar{r}^{a,\,n_e}$ for EnKF method a and $\bar{r}^{b,\,n_e}$ for EnKF method b, where $\bar{r}^{a,\,n_e} < \bar{r}^{b,\,n_e}$ 55, 56

$\mathbf{d}_i \in \mathbb{R}^{n_m}$

member of the ensemble of observations in the EnKF 21, 48–50, 75, 78

d_-^{max}

largest insignificant relative difference $d^{a<b,\,n_e}$ of RMSE means of a method comparison 56

d_+^{min}

smallest significant relative difference $d^{a<b,\,n_e}$ of RMSE means of a method comparison 56

$D_-^{n_e}$

set of insignificant relative differences $d^{a<b,\,n_e}$ of RMSE means of a method comparison 56

$D_+^{n_e}$

set of significant relative differences $d^{a<b,\,n_e}$ of RMSE means of a method comparison 56

dV $[L^3]$

control volume in the definition of the source term W of the groundwater flow equation 11

F

cumulative distribution function of a predicted state vector element in the normal score EnKF 50, 51

G

cumulative distribution function of the Gaussian pdf with zero mean and unit standard deviation used in the normal score EnKF 50

G

Gaussian bell curve in the local EnKF 49

g $[LT^{-2}]$

gravitational constant in the groundwater flow equation 10, 11, 13

$\mathbf{H} \in \mathbb{R}^{n_m \times (n_p + n_d)}$

measurement operator of the pilot point Kalman filter and the PP-EnKF 77–80

$\mathbf{H} \in \mathbb{R}^{n_m \times n_s}$

measurement operator of the Kalman filter and the EnKF 18, 19, 21, 48, 75

h [m]

water table drawdown in the Theis model 33

h [L]

reference hydraulic head in the groundwater flow equation 10, 12–15

h_0 [m]

unperturbed water table in the Theis model 33

h_{obs} [L]

random variable for an observed hydraulic head 84

$\mathbf{H}_p \in \mathbb{R}^{n_m \times n_{s,p}}$

measurement operator projected onto the parameter space in the dual EnKF 49

$\mathbf{H}_s \in \mathbb{R}^{n_m \times n_{s,s}}$

measurement operator projected onto the space of dynamic variables in the dual EnKF 50

$\mathbf{K} \in \mathbb{R}^{n_s \times n_m}$

Kalman gain matrix, computed from (1) covariance matrices in the Kalman filter and (2) sample covariance matrices in the EnKF 19, 21, 48

$\mathbf{K}$ [L^2]

hydraulic permeability tensor in the groundwater flow equation 10, 11, 13

k [$\mathrm{m\,s^{-1}}$]

hydraulic conductivity of the Theis problem 33

$\mathbf{k}$ [LT^{-1}]

hydraulic conductivity tensor in the groundwater flow equation 13

$\mathbf{K}_{loc} \in \mathbb{R}^{n_s \times n_m}$

Kalman gain after elementwise multiplication with correlation matrix in the local EnKF 48

$K_{\log}$ [−]

random variable for an isotropic logarithmic hydraulic permeability value 11, 84

$\mathbf{K}_p \in \mathbb{R}^{n_{s,p} \times n_m}$

Kalman gain matrix projected onto the parameter space in the dual EnKF 49

$\mathbf{K}_s \in \mathbb{R}^{n_{s,s} \times n_m}$

Kalman gain matrix projected onto the space of dynamic variables in the dual EnKF 50

k_x $[LT^{-1}]$

hydraulic conductivity in x-direction in the groundwater flow equation 12–14

$k_x^{(-)}$ $[LT^{-1}]$

hydraulic conductivity in x-direction at cell boundary in negative x-direction in the discretization of the groundwater flow equation 13–15

$k_x^{(+)}$ $[LT^{-1}]$

hydraulic conductivity in x-direction at cell boundary in positive x-direction in the discretization of the groundwater flow equation 13–15

k_y $[LT^{-1}]$

hydraulic conductivity in y-direction in the groundwater flow equation 13, 14

$k_y^{(-)}$ $[LT^{-1}]$

hydraulic conductivity in y-direction at cell boundary in negative y-direction in the discretization of the groundwater flow equation 13–15

$k_y^{(+)}$ $[LT^{-1}]$

hydraulic conductivity in y-direction at cell boundary in positive y-direction in the discretization of the groundwater flow equation 13–15

k_z $[LT^{-1}]$

hydraulic conductivity in z-direction in the groundwater flow equation 13, 14

$k_z^{(-)}$ $[LT^{-1}]$

hydraulic conductivity in z-direction at cell boundary in negative z-direction in the discretization of the groundwater flow equation 13–15

$k_z^{(+)}$ $[LT^{-1}]$

hydraulic conductivity in z-direction at cell boundary in positive z-direction in the discretization of the groundwater flow equation 13–15

L

dimension symbol for length 10–13, 48, 84

M

forward solution of the ensemble Kalman filter 20, 75

M

dimension symbol for mass 10–12

$\mathbf{M} \in \mathbb{R}^{n_s \times n_s}$

matrix for linear forward computation of the Kalman filter 18, 74, 75

$\mathbb{N}$

set of natural numbers 17, 18, 20, 21, 49, 55, 75, 77

N

dimension symbol for amount of substance 84

$n_d \in \mathbb{N}$

number of dynamic variables in the pilot point Kalman filter 77

$n_e \in \mathbb{N}$

ensemble size of the ensemble Kalman filter 20, 21, 50, 55, 62, 75, 78, 83

$n_g \in \mathbb{N}$

number of grid cells of a model domain 53, 55, 83

$n_m \in \mathbb{N}$

number of observations in the Kalman filter or EnKF 18, 21, 75

$n_p \in \mathbb{N}$

number of pilot point parameters in the pilot point Kalman filter 77

$n_r \in \mathbb{N}$

number of non-pilot point parameters in the pilot point Kalman filter 77

NS

normal score transform used in the normal score EnKF 50, 51

$n_s \in \mathbb{N}$

size of the state vector of the Kalman filter or EnKF 17, 20, 49, 74, 77

$n_{s,p} \in \mathbb{N}$

number of parameters in the state vector in the dual EnKF 49

$n_{s,s} \in \mathbb{N}$

number of state variables in the state vector in the dual EnKF 49

$n_{syn} \in \mathbb{N}$

number of synthetic experiments in a method comparison 55, 58, 62, 64

$n_t \in \mathbb{N}$

number of assimilation times of the Kalman filter or ensemble Kalman filter 20, 51

$\mathbf{P} \in \mathbb{R}^{n_s \times n_s}$

covariance matrix of the predicted state vector of the Kalman filter, result of the forward computation 18, 19, 75, 77–79

$P\ [ML^{-1}T^{-2}]$

fluid pressure in the definition of the reference hydraulic head h 10

$\mathbf{P}^0 \in \mathbb{R}^{n_s \times n_s}$

initial covariance matrix of the Kalman filter 18, 19, 75

$\mathbf{P}^a \in \mathbb{R}^{n_s \times n_s}$

covariance matrix of the updated state vector of the Kalman filter 19, 75, 78

$p^{a<b,\,n_e}$

fraction of synthetic experiments, for which one EnKF method a yields a smaller RMSE than another EnKF method b 55, 56

$\mathbf{P}^a_{dd} \in \mathbb{R}^{n_d \times n_d}$

updated covariances between pairs of dynamic variables in the pilot point Kalman filter 77, 78

$\mathbf{P}^a_{dp} \in \mathbb{R}^{n_d \times n_p}$

updated covariances between dynamic variables and parameters at pilot points in the pilot point Kalman filter 77, 78

$\mathbf{P}_{pd}^{a} \in \mathbb{R}^{n_p \times n_d}$

updated covariances between parameters at pilot points and dynamic variables in the pilot point Kalman filter 77, 78

$\mathbf{P}_{pp}^{a} \in \mathbb{R}^{n_p \times n_p}$

updated covariances between all pairs of parameters at pilot points in the pilot point Kalman filter 77, 78

$p^{b<a,\, n_e}$

fraction of synthetic experiments, for which one EnKF method b yields a smaller RMSE than another EnKF method a 55

$\mathbf{P}_{dd} \in \mathbb{R}^{n_d \times n_d}$

covariances between pairs of dynamic variables in the pilot point Kalman filter 77–79

$\mathbf{P}_{dd,e} \in \mathbb{R}^{n_d \times n_d}$

sample covariances between pairs of dynamic variables in the pilot point EnKF 78–80

$\mathbf{P}_{dp} \in \mathbb{R}^{n_d \times n_p}$

covariances between dynamic variables and parameters at pilot points in the pilot point Kalman filter 77–79

$\mathbf{P}_{dp,e} \in \mathbb{R}^{n_d \times n_p}$

sample covariances between dynamic variables and parameters at pilot points in the pilot point EnKF 78–80

$\mathbf{P}_{dr} \in \mathbb{R}^{n_d \times n_r}$

covariances between dynamic variables and parameters at non-pilot points in the pilot point Kalman filter 77, 79

$\mathbf{P}_{dr,e} \in \mathbb{R}^{n_d \times n_r}$

sample covariances between dynamic variables and parameters at non-pilot points in the pilot point EnKF 80

$\mathbf{P}_{e} \in \mathbb{R}^{n_s \times n_s}$

ensemble covariance matrix of states and parameters of the ensemble Kalman filter 21, 48, 49, 75, 80

$\mathbf{P}_{\text{hybrid}} \in \mathbb{R}^{n_s \times n_s}$

covariance matrix of the hybrid EnKF 49

$P_p \in \mathbb{R}^{n_s \times (n_p + n_d)}$

interpolation operator in the pilot point Kalman filter and the PP-EnKF 78–80

$\mathbf{P}_{pd} \in \mathbb{R}^{n_p \times n_d}$

covariances between parameters at pilot points and dynamic variables in the pilot point Kalman filter 77–79

$\mathbf{P}_{pd,e} \in \mathbb{R}^{n_p \times n_d}$

sample covariances between parameters at pilot points and dynamic variables in the pilot point EnKF 78–80

$\mathbf{P}_{pp} \in \mathbb{R}^{n_p \times n_p}$

covariances between all pairs of parameters at pilot points in the pilot point Kalman filter 77–79

$\mathbf{P}_{pp,e} \in \mathbb{R}^{n_p \times n_p}$

sample covariances between all pairs of parameters at pilot points in the pilot point EnKF 78–80

$\mathbf{P}_{pr} \in \mathbb{R}^{n_p \times n_r}$

covariances between parameters at pilot points and parameters at non-pilot points in the pilot point Kalman filter 77, 79

$\mathbf{P}_{pr,e} \in \mathbb{R}^{n_p \times n_r}$

sample covariances between parameters at pilot points and parameters at non-pilot points in the pilot point EnKF 80

$\mathbf{P}_{rd} \in \mathbb{R}^{n_r \times n_d}$

covariances between parameters at non-pilot points and dynamic variables in the pilot point Kalman filter 77, 79

$\mathbf{P}_{rd,e} \in \mathbb{R}^{n_r \times n_d}$

sample covariances between parameters at non-pilot points and dynamic variables in the pilot point EnKF 80

$\mathbf{P}_{rp} \in \mathbb{R}^{n_r \times n_p}$

covariances between parameters at non-pilot points and parameters at pilot points in the pilot point Kalman filter 77, 79

$\mathbf{P}_{rp}^0 \in \mathbb{R}^{n_r \times n_p}$

fixed covariance matrix between pilot point parameters and non-pilot point parameters in the pilot point Kalman filter and the PP-EnKF 78–80, 82

$\mathbf{P}_{rp,e} \in \mathbb{R}^{n_r \times n_p}$

sample covariances between parameters at non-pilot points and parameters at pilot points in the pilot point EnKF 80

$\mathbf{P}_{rr} \in \mathbb{R}^{n_r \times n_r}$

covariances between pairs of parameters at non-pilot points in the pilot point Kalman filter 77, 79

$\mathbf{P}_{rr,e} \in \mathbb{R}^{n_r \times n_r}$

sample covariances between pairs of parameters at non-pilot points in the pilot point EnKF 80

$\mathbf{P}_{\text{static}} \in \mathbb{R}^{n_s \times n_s}$

static component of the covariance matrix of the hybrid EnKF 49

$\mathbf{P}_{yppy} \in \mathbb{R}^{n_m \times n_m}$

covariance matrix among simulated measurements in the pilot point Kalman filter 77, 79

$\mathbf{P}_{yppy,e} \in \mathbb{R}^{n_m \times n_m}$

sample covariance matrix among simulated measurements in the pilot point EnKF 78, 79

$\mathbf{P}_{yy} \in \mathbb{R}^{n_m \times n_m}$

covariance matrix among simulated measurements in the Kalman filter 79

$\mathbf{P}_{yy,e} \in \mathbb{R}^{n_m \times n_m}$

sample covariance matrix among simulated measurements in the EnKF 79

Q [$\mathrm{m}^3\,\mathrm{s}^{-1}$]

pumping rate of the well in the Theis model 33

$q^{a<b,\,n_e} \in [0,1]$

quotients of RMSE means $\bar{r}^{a,\,n_e}$ from synthetic experiments with ensemble size n_e for EnKF methods a and b 55

$\mathbb{R}$

set of real numbers 50, 51

$\mathbf{R} \in \mathbb{R}^{n_m \times n_m}$

measurement error matrix of the Kalman filter and the EnKF 18, 19, 21, 48, 75, 77, 78

r [m]

distance from the well in the Theis model 33

$\bar{r}^{a,\,n_e} \in \mathbb{R}^+$

RMSE mean for given EnKF method a and ensemble size n_e 55, 56

$\bar{r}_X^{a,\,n_e} \in \mathbb{R}^+$

RMSE mean for given set of synthetic experiments X, EnKF method a and ensemble size n_e 55

$\bar{r}^{b,\,n_e} \in \mathbb{R}^+$

RMSE mean for given EnKF method b and ensemble size n_e 55

$\bar{r}_X^{b,\,n_e} \in \mathbb{R}^+$

RMSE mean for given set of synthetic experiments X, EnKF method b and ensemble size n_e 55

$r_j^{a,\,n_e} \in \mathbb{R}^+$

single RMSE for given synthetic experiment j, EnKF method a and ensemble size n_e, same as RMSE_j, but inputs made explicit 55

$\mathrm{RMSE}_j \in \mathbb{R}^+$

root-mean-square error of a single synthetic experiment j with given EnKF method and ensemble size 53, 83

$\mathbb{R}^+$

set of positive real numbers 48, 51, 52, 55

S [$-$]

storage coefficient of the Theis problem 33

S_s [L^{-1}]

specific storage coefficient in the groundwater flow equation 10, 13–15, 33

$\mathrm{STD}_j \in \mathbb{R}^+$

overall standard deviation of a single synthetic experiment j with given ensemble size for one of the EnKF methods in the synthetic experiment 83

$T \in \mathbb{R}^+$ [T]

computing time of a given EnKF run 51

T

dimension symbol for time 10–13, 51

T [Θ]

temperature in the heat transport equation 11

t [s]

the time variable in the Theis problem 33

t [T]

the time variable in the groundwater flow equation or other transient simulations 10, 11, 13, 20

$T_{\text{Dual}} \in \mathbb{R}^+$ [T]

computing time of a given run of the dual EnKF 51

$T_{\text{Iterative}} \in \mathbb{R}^+$ [T]

computing time of a given run of the iterative EnKF 51

$\mathbf{v}$ [LT^{-1}]

Darcy specific discharge in the groundwater flow equation 10, 11

W [T^{-1}]

sources and sinks term of the groundwater flow equation 10, 11, 13

w

well function used in the Theis model 33

W_V [L^3T^{-1}]

external volume flow in the definition of the sources and sinks term W in the groundwater flow equation 11

X

subset of synthetic experiments of a method comparison 55

$\mathbf{x} \in \mathbb{R}^{n_s}$

State vector of the Kalman filter 17

$\mathbf{x}^0 \in \mathbb{R}^{n_s}$

initial mean state vector of the Kalman filter 18, 19, 74

$\mathbf{x}_i^0 \in \mathbb{R}^{n_s}$

member of the ensemble of initial states of the EnKF 75

$\mathbf{x}^a \in \mathbb{R}^{n_s}$

updated state vector of the Kalman filter, result of the analysis/update equation 18, 19, 75, 78

$\mathbf{x}_d^a \in \mathbb{R}^{n_d}$

dynamic variables in the updated state vector of the pilot point Kalman filter 77, 78

$\mathbf{x}_{d,i}^a \in \mathbb{R}^{n_d}$

dynamic states part of the state vector after the state update in the pilot point EnKF 78

$x_i^a \in \mathbb{R}$

single component of a member of the ensemble of updated states of the EnKF, result of the analysis/update equation 50, 51

$\mathbf{x}_i^a \in \mathbb{R}^{n_s}$

member of the ensemble of updated states of the EnKF, result of the analysis/update equation 20, 21, 48, 75, 78

$\mathbf{x}_p^a \in \mathbb{R}^{n_p}$

pilot point parameters in the updated state vector of the pilot point Kalman filter 77, 78

$\mathbf{x}_{p,i}^a \in \mathbb{R}^{n_p}$

pilot point parameter part of the updated state vector in the pilot point EnKF 78

$\mathbf{x}_{p,i}^a \in \mathbb{R}^{n_{s,p}}$

parameter part of the state vector after the parameter update in the dual EnKF 49, 50

$\mathbf{x}_{s,i}^a \in \mathbb{R}^{n_{s,s}}$

state variable part of the state vector after the state update in the dual EnKF 50

$\mathbf{x}^f \in \mathbb{R}^{n_s}$

predicted state vector of the Kalman filter, result of the forward computation 18, 19, 74, 75, 77, 78

$\mathbf{x}_d^f \in \mathbb{R}^{n_d}$

dynamic variables in the predicted state vector of the pilot point Kalman filter 77, 78

$\mathbf{x}_{d,i}^f \in \mathbb{R}^{n_d}$

dynamic states part of the state vector in the pilot point EnKF 78

$x_i^f \in \mathbb{R}$

single component of a member of the ensemble of predicted states of the EnKF 50, 51

$\mathbf{x}_i^f \in \mathbb{R}^{n_s}$

member of the ensemble of predicted states of the EnKF, result of the forward computation 20, 21, 48–50, 75, 78

$\mathbf{x}_p^f \in \mathbb{R}^{n_p}$

pilot point parameters in the predicted state vector of the pilot point Kalman filter 77, 78

$\mathbf{x}_{p,i}^f \in \mathbb{R}^{n_p}$

pilot point parameter part of the state vector in the pilot point EnKF 78

$\mathbf{x}_{p,i}^f \in \mathbb{R}^{n_{s,p}}$

parameter part of the state vector in the dual EnKF 49

$\mathbf{x}_r^f \in \mathbb{R}^{n_r}$

non-pilot point parameters in the predicted state vector of the pilot point Kalman filter 77

$\mathbf{x}_{r,i}^f \in \mathbb{R}^{n_r}$

non-pilot point parameter part of the state vector in the pilot point EnKF 78

$\mathbf{x}_{s,i}^f \in \mathbb{R}^{n_{s,s}}$

state variable part of the state vector after the first forward computation in the dual EnKF 49

$\mathbf{x}_{s,i}^{f,2} \in \mathbb{R}^{n_{s,s}}$

state variable part of the state vector after the second forward computation in the dual EnKF 50

$\mathbf{x}^{\text{meas}} \in \mathbb{R}^{n_s}$

hypothetical true state vector that would lead to measurement observations after applying the measurement operator H in the EnKF 21

$x_0^s \in \mathbb{R}$

left artificial support point for the inverse normal score transform in the normal score EnKF 51

$x_1^s \in \mathbb{R}$

right artificial support point for the inverse normal score transform in the normal score EnKF 51

$\mathbf{y} \in \mathbb{R}^{n_m}$

means of measurement observations before perturbation in the EnKF 21

$\bar{\mathbf{Y}}_j \in \mathbb{R}^{n_g}$

vector containing the estimated mean logarithmic permeabilities across a model domain for synthetic experiment j 55, 83

$\bar{Y}_{ij} \in \mathbb{R}$

estimated mean logarithmic permeability at grid cell i for synthetic experiment j 53, 83

$Y_{k,i,j} \in \mathbb{R}$

logarithmic permeability of realization k at grid cell i for synthetic experiment j 83

$\mathbf{Y}_j^t \in \mathbb{R}^{n_g}$

vector containing the synthetic reference logarithmic permeabilities across a model domain for synthetic experiment j 55, 83

$Y_{ij}^t \in \mathbb{R}$

synthetic reference logarithmic permeability at grid cell i for synthetic experiment j 53, 83

z $[L]$

height above a reference height in the groundwater flow equation 10, 13

Δt $[T]$

discretization time interval 13–15

Δx $[L]$

discretization interval in x-direction 13–15

$\Delta x_i^{(-)}$ $[L]$

distance between the centers of neighboring grid cells of cell i in negative x-direction 14, 15

$\Delta x_i^{(+)}$ $[L]$

distance between the centers of neighboring grid cells of cell i in positive x-direction 14, 15

Δy $[L]$

discretization interval in y-direction 13–15

$\Delta y_j^{(-)}$ $[L]$

distance between the centers of neighboring grid cells of cell j in negative y-direction 14, 15

$\Delta y_j^{(+)}$ $[L]$

distance between the centers of neighboring grid cells of cell j in positive y-direction 14, 15

Δz [m]

thickness of the aquifer in the Theis model 33

Δz $[L]$

discretization interval in z-direction 13–15

$\Delta z_k^{(-)}$ $[L]$

distance between the centers of neighboring grid cells of cell k in negative z-direction 14, 15

$\Delta z_k^{(+)}$ $[L]$

distance between the centers of neighboring grid cells of cell k in positive z-direction 14, 15

Θ

dimension symbol for thermodynamic temperature 11, 12

$\alpha \in (0, 1]$

damping factor of the damped EnKF 48, 53

$\beta \in [0, 1]$

covariance weight of the hybrid EnKF 49, 53, 64, 67

$\epsilon_i \in \mathbb{R}^{n_m}$

perturbation ensemble for observations generated from Gaussian noise 21

$\lambda \in \mathbb{R}^+$ $[L]$

characteristic length scale of the correlation matrix in the local EnKF 48, 49, 53, 64, 67

λ $[MLT^{-3}\Theta^{-1}]$

thermal conductivity tensor of a matrix-fluid mixture in the heat transport equation 11

μ_f $[ML^{-1}T^{-1}]$

fluid dynamic viscosity in the groundwater flow equation 10, 11, 13

ρ

spherical correlation function used for the simulation of permeability fields 52

$\rho \in \mathbb{R}^{n_s \times n_m}$

correlation matrix in the local EnKF 48, 49

ρ_0 $[ML^{-3}]$

reference fluid density of the reference hydraulic head from the groundwater flow equation 10, 11, 14

ρ^* $[ML^{-3}]$

source fluid density in the definition of the source term W 11

$\rho(c_{obs}, K_{\log})$

correlation function between tracer concentration at an observation location and the logarithmic permeability at a grid location 84

ρ_f $[ML^{-3}]$

fluid density in the groundwater flow equation 10–13

$\rho(h_{obs}, K_{\log})$

correlation function between hydraulic head at an observation location and the logarithmic permeability at a grid location 84

ρ_m $[ML^{-3}]$

matrix density in the heat transport equation 11, 12

ρ_r $[-]$

relative fluid density variation with respect to a reference density ρ_0 10, 13, 14

$\rho_r^{(-)}$ $[-]$

relative fluid density variation at cell boundary in negative z-direction 14, 15

$\rho_r^{(+)}$ $[-]$

relative fluid density variation at cell boundary in positive z-direction 14, 15

$\sigma_{\mathtt{c}} \in \mathbb{R}^+$ $[N/L^3]$

Noise of concentration measurements 53, 82, 84

$\sigma_{\mathtt{h}} \in \mathbb{R}^+$ $[L]$

Noise of hydraulic head measurements 53, 82, 84

σ_{ij}^2

sample variance of the logarithmic permeability at grid cell i of synthetic experiment j 83

$\sigma_{\mathtt{K}} \in \mathbb{R}^+$ $[\log(1/L^2)]$

sample standard deviation of logarithmic permeability values of a single isotropic permeability 84

φ $[-]$

porosity in the groundwater flow equation 10, 11

List of acronyms and definitions

AD

Automatic Differentiation (AD) is a technique to evaluate the derivatives of a function defined by a computer program[1] 3, 27

ASCII

The American Standard Code for Information Interchange (ASCII) is a character encoding standard for electronic communication[2] 29, 31

Bi-CGStab

The Bi-Conjugate Gradients Stabilized (Bi-CGStab) method is a fast and smoothly converging iterative variant of the Bi-Conjugate Gradients (Bi-CG) method for the numerical solution of nonsymmetric linear systems (van der Vorst, 1992) 4, 24, 33, 52

BLAS

The Basic Linear Algebra Subprograms (BLAS) are Fortran-callable routines that provide standard building blocks for performing basic vector and matrix operations (Lawson et al., 1979, Dongarra et al., 1988, 1990) 4, 25

CMake

CMake is an open-source, cross-platform family of tools designed to build, test and package software[3] 29

CPU

The central processing unit (CPU) is the electronic circuitry within a computer that executes instructions that make up a computer program[4] 42

damped EnKF

The damped EnKF is an EnKF method that reduces the filter inbreeding problem with the help of a damping parameter (Hendricks Franssen and Kinzelbach, 2008) 6, 32, 53, 56, 58, 60, 62, 64, 69, 72, 82, 86, 90, 92, 95

[1] https://www-sop.inria.fr/tropics/ad/whatisad.html [Accessed 2020/04/26]
[2] https://en.wikipedia.org/wiki/ASCII [Accessed 2020/05/11]
[3] https://cmake.org/ [Accessed 2020/05/11]
[4] https://en.wikipedia.org/wiki/Central_processing_unit [Accessed 2020/05/11]

Doxygen

Doxygen is the de facto standard tool for generating documentation from annotated C++ sources, but it also supports other popular programming languages[5] 27, 28, 30

dual EnKF

The dual EnKF is an EnKF method for sequential estimation of both parameters and state variables of a hydrologic model by separating the update equation for parameters and state variables 6, 25, 32, 46, 49, 51, 56, 58, 60, 62, 69, 90, 92

DuMuX

DUNE for multi-{phase, component, scale, physics,...} (DuMuX) is a free and open-source simulator for flow and transport processes in porous media, based on the Distributed and Unified Numerics Environment (DUNE) (Flemisch et al., 2011) 3

EFCOSS

The *Environment for Combining Optimization and Simulation Software* (EFCOSS) is an interactive software environment that combines numerical simulation codes with optimization software packages in an automated and modular way (Rasch and Bücker, 2010) 5

EnKF

The ensemble Kalman filter (EnKF) is a sequential Monte Carlo method that provides an alternative to the traditional Kalman filter to better handle large state spaces and nonlinear error evolution (see Evensen, 2009) III–V, VII–IX, 1–7, 9, 17, 19–21, 23, 25, 27, 30–32, 37, 39, 40, 45–48, 50, 51, 53, 55, 56, 58–62, 64, 66–76, 79–84, 86, 90–94, 97–99, 101, 103

EoCoE

The Energy oriented Centre of Excellence (EoCoE) is part of the Horizon 2020 European funding for Research & Innovation[6] 42

FEFLOW

The finite element subsurface FLOW simulation system (FEFLOW) solves the flow, mass, and heat transport equations in porous and fractured media by a multidimensional finite element method for complex geometric and parametric situations (Diersch, 2014) 3

Fortran

Fortran is a general-purpose, compiled imperative programming language that is especially suited to numeric computation and scientific computing[7] 4, 25, 31

[5] http://doxygen.nl/ [Accessed 2020/05/11]
[6] https://www.eocoe.eu/ [Accessed 2020/05/11]
[7] https://en.wikipedia.org/wiki/Fortran [Accessed 2020/04/26]

GSLib

The Geostatistical Software Library (GSLib) was originally used for a collection of geostatistical programs developed at Stanford University (Deutsch and Journel, 1995) 4, 25, 31

HDF5

The Hierarchical Data Format (HDF5) is a data model, library, and file format for storing and managing data[8] 4, 25, 29, 33, 35

hybrid EnKF

The hybrid EnKF is an EnKF method that operates by perturbed observations that are generated by adding random noise consistent with observation error statistics to the control set of observations (Hamill and Snyder, 2000) 6, 25, 32, 46, 49, 53, 56, 58, 60, 62, 64, 67, 69, 71, 72, 74, 76, 80, 82, 94, 95, 98

iterative EnKF

The iterative EnKF is an iterative formulation of the EnKF for strongly nonlinear systems in the perfect-model framework (Sakov et al., 2012) 6, 25, 32, 46, 51, 56, 58, 60, 62, 69, 72, 94, 95

kriging

Kriging is a method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior covariances[9] 7, 90

LAPACK

The Linear Algebra PACKage (LAPACK) is a library of Fortran 77 subroutines for solving the most commonly occurring problems in numerical linear algebra (Anderson et al., 1999) 4, 25

local EnKF

The local EnKF is an EnKF method that reduces noisiness by employing a distance-dependent reduction of background error covariances estimates (Hamill et al., 2001) 6, 25, 32, 46, 53, 56, 58, 60, 62, 64, 67, 69, 71, 72, 74, 76, 80–82, 98

master

Git-branch of SHEMAT-Suite optimized for pure forward computation 28, 29, 43

[8] https://portal.hdfgroup.org/display/HDF5/HDF5 [Accessed 2020/04/26]
[9] https://en.wikipedia.org/wiki/Kriging [Accessed 2020/04/26]

master-ad

Git-branch of SHEMAT-Suite for AD-mode, deterministic inversion 28, 30

master-all

Git-branch of SHEMAT-Suite containing the full functionality of the software 28, 43

master-sm

Git-branch of SHEMAT-Suite for stochastic mode, deterministic inversion 28, 30, 31

MIT license

The MIT License is a permissive free software license originating at the Massachusetts Institute of Technology (MIT) in the late 1980s[10] 102

MODFLOW

The USGS Modular Three-Dimensional Finite-Difference Ground-Water Flow Model, or MODFLOW, is a popular open-source groundwater flow model distributed by the U.S. Geological Survey (Hughes et al., 2017) 3

Monte Carlo

Monte Carlo (MC) is a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results[11] III, VII, 3, 4, 7, 19, 25, 27, 28, 42, 47, 73

MPI

The Message Passing Interface (MPI) is a standardized and portable message-passing standard designed by a group of researchers from academia and industry to function on a wide variety of parallel computing architectures[12] 42

normal score EnKF

The normal score EnKF is an EnKF method that ensures that the joint state vector of the EnKF follows marginal Gaussian distributions at all locations and all times by applying a normal-score transform (Zhou et al., 2011) 6, 25, 32, 46, 56, 58, 60, 62, 69

OpenGeoSys

OpenGeoSys is a scientific open-source initiative for numerical simulation of thermo-hydro-mechanical-chemical processes in porous media (Kolditz et al., 2012) 3

[10] https://en.wikipedia.org/wiki/MIT_License [Accessed 2020/05/11]
[11] https://en.wikipedia.org/wiki/Monte_Carlo_method [Accessed 2020/04/26]
[12] https://en.wikipedia.org/wiki/Message_Passing_Interface [Accessed 2020/05/11]

OpenMP

OpenMP is a specification for a set of compiler directives, library routines, and environment variables that can be used to specify high-level parallelism in Fortran and C/C++ programs[13] 3, 26, 29

PDAF

The Parallel Data Assimilation Framework (PDAF) is a software environment for ensemble data assimilation[14] 42

pdf

probability density function 50, 70

PDI

The portable data interface (PDI) offers to exchange data between the application code and various external data handlers, such as for example the file-system for I/O or another code for code-coupling[15] 42

PETSc

The Portable, Extensible Toolkit for Scientific Computation (PETSc) is a suite of data structures and routines for the scalable (parallel) solution of scientific applications modeled by partial differential equations[16] 42

PP-EnKF

The pilot point ensemble Kalman filter (PP-EnKF) is an EnKF method tailored towards reducing spurious correlations in parameter estimation by updating in a first step only parameters at a fixed subset of locations, called pilot points (Chapter 5) IV, V, VIII, IX, 7, 8, 71, 73, 74, 76, 78–84, 86, 90, 92, 94, 95, 97–99, 101, 103, 104

RMSE

Root-Mean-Square Error IV, VIII, 5, 7, 40, 47, 53, 55–58, 60–62, 64, 65, 67–70, 73, 74, 83–86, 90, 92, 94, 95, 97–99, 103, 104

SGSim

Sequential Gaussian Simulation (SGSim) is part of the GSLIB 4, 25, 31, 39, 53, 82

[13] https://www.openmp.org/about/openmp-faq/#WhatIs [Accessed 2020/04/26]
[14] http://pdaf.awi.de/trac/wiki [Accessed 2020/05/11]
[15] https://pdi.julien-bigot.fr/0.6.0/Concepts.html [Accessed 2020/05/11]
[16] https://www.mcs.anl.gov/petsc/ [Accessed 2020/05/11]

SHEMAT-Suite

The Simulator for HEat and MAss Transport suite (SHEMAT-Suite) is a numerical software for simulating flow, heat and species transport enhanced with various capabilities for inverse problems and data assimilation III, V, VII–IX, 1, 3–5, 7, 9, 11–16, 21–29, 31–33, 35–37, 39, 40, 42, 43, 81, 101–104

spurious correlations

Correlation bias in EnKF methods caused by small ensemble sizes III–V, 1, 6, 7, 71, 73, 74, 76, 79, 80, 83, 86, 95, 101, 104

STOMP

Subsurface Transport Over Multiple Phases (STOMP) is a scalable implementation of a multifluid subsurface flow and transport simulator with capabilities for volatile organic compounds (White et al., 2008) 3

synthetic experiment

A comparison of a number of data assimilation methods for a given groundwater flow problem and for a given set of synthetically generated observations IV, 5–8, 45, 47, 48, 51, 53, 55, 58, 62, 64, 65, 67, 69, 70, 73, 74, 82–84, 86, 89, 92, 94, 96–99, 103, 104

Tapenade

Tapenade is an Automatic Differentiation (AD) tool which, given a Fortran or C code that computes a function, creates a new code that computes its tangent or adjoint derivatives (Hascoët and Pascual, 2013) 4, 25

TOUGH

The *Transport Of Unsaturated Groundwater and Heat* suite (TOUGH) are multi-dimensional numerical models for simulating the coupled transport of water, vapor, non-condensible gas, and heat in porous and fractured media (Finsterle et al., 2014) 3

uindex

Unit index of SHEMAT-Suite. One unit corresponds to a set of input parameters 35–37

USGS

United States Geological Survey 4

VTK

The Visualization Toolkit (VTK) is an open-source, freely available software system for 3D computer graphics, modeling, image processing, volume rendering, scientific visualization, and 2D plotting (Schroeder et al., 2006) 4, 25, 33

Acknowledgments

Resources This thesis was supported by the Deutsche Forschungsgemeinschaft. Simulations were performed with computing resources granted by RWTH Aachen University under project rwth0009.

Personal First and foremost, I wish to thank my main supervisor Harrie-Jan Hendricks Franssen for the scientific guidance, in particular the numerous revisions of my manuscripts that he had to compile.

Thanks to Christoph Clauser for his supervision and guidance during the *Doktorandengespräche*.

Thanks to Gabriele Marquart and Norbert Klitzsch for scientific supervision of the chapters about the EnKF method comparison and the SHEMAT-Suite software, respectively.

I would like to express thanks to the PhD committee members Julia Kowalski, Florian Wagner and Wolfgang Nowak. I feel additional gratitude for Wolfgang Nowak for working together on the pilot point ensemble Kalman filter.

Thanks to the remaining co-authors of the chapter on SHEMAT-Suite: Volker Rath, Johanna Fink, Darius Mottaghy, Andreas Wolf, Ralf Seidler, H. Martin Bücker. Working on the manuscript has been a pleasure. Special thanks to Johanna Fink for letting me use the results of her inverse simulation in Section 3.3.2.

Thanks to Christian Vogt for the work on ensemble Kalman filter that was the foundation of this thesis.

Thanks to my room-mates for the good working environment!

- Aachen I: Saurabh, Chris
- Jülich I: Wei, Betiglu, Sebastian, Markus, Emil
- Aachen II: Thomas, Wei, Lucas, Nour, Lothar, Ariel, Mathis
- Jülich II: Shehan, Xiaoran, Sibghat

Thanks to all kind colleagues who have become friends, especially Johanna O., Paromita, Jan, Henrik, Tao, Alexander, Sandra.

Thanks to my friends and family for support and motivation.

Thanks to Christin for starting this PhD with me as my girlfriend and even more thanks for ending it with me as my wife and mother of our child. In-between, you became my co-author and showed me how to finish a PhD in a reasonable amount of time. You are the best companion I could wish for.