Estimating Hydraulic Anisotropy in Stratified Aquifers

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Ruth Verena Maier
aus Klagenfurt
Österreich

Tübingen
2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation	19.11.2021
Stellvertretender Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter	Prof. Dr.-Ing. Olaf A. Cirpka
2. Berichterstatter	Prof. Dr. Warren Barrash

Acknowledgements

First and foremost I want to thank my supervisors Olaf A. Cirpka and Carsten Leven-Pfister for giving me the opportunity to be part of this exciting project and for their guidance during the journey towards my PhD. Thank you Olaf for giving me direction, for the many insightful discussions we had during the process and, most of all, for your constructive and invaluable feedback throughout. I have learnt so much from you and working with you has undoubtedly made me a better researcher. Thank you Carsten for your guidance throughout the research and writing stages and for your support in relation to just about anything. Working in the field with you was a privilege during which I did not only learn a lot but which, together with your valuble feedback and our discussions, has inspired me to really engage with my research.

A special thank you also goes to Prof. Dr. Warren Barrash for showing such an interest in my work and for agreeing to examine my thesis and to become part of my PhD committee.

I also want to thank Dr.-Ing. Bernhard Odenwald, head of the Groundwater Department at the Federal Waterways Engineering and Research Institute, for making this project possible.

A special thank you to Daniel Strasser for your support and feedback during my thesis. Performing the field tests without you wouldn't have been possible. I also want to thank you for our prolonged and insightful discussions from which I have always learned a lot.

Thank you also to Philipp Schäfer, for your practical and logistical support regarding the installation of the well network at the field site Kappel-Grafenhausen.

A special thanks to Paul Deißinger, Malte Kleemann and Maximilian Stoll for the collaboration. The opportunity to supervise your work was an enriching experience and I am thankful for your many valuable contributions to this project.

Thank you Emilio for your support throughout my PhD, for your constructive and valuable feedback and for our many lively and extensive discussions on data processing and data inversion.

Last but certainly not least I want to thank my whole family for all the support throughout my thesis. A special thanks to my Dad for giving me the final push and having faith in me that I am able to finish this work.

Abstract

Understanding groundwater flow processes requires detailed information on the heterogeneity and anisotropy of hydraulic conductivity. In sedimentary aquifers, the spatial variability of hydraulic conductivity results from the stratification of material and primarily occurs between individual horizontal layers. On large scales, the vertical heterogeneity induces hydraulic anisotropy with the horizontal conductivity typically exceeding the vertical one. Despite the importance of hydraulic anisotropy, for example in the design of remediation systems or in the planning of dewatering measures in construction pits or mines, research focusing on estimating the difference between the horizontal and vertical hydraulic conductivity has so far been limited.

This work investigates the potential of a field method for estimating the large-scale hydraulic anisotropy induced by vertical heterogeneity in stratified aquifers. The approach is based on the inversion of drawdown data steming from pumping tests in a partially penetrating well in which water is successively extracted from different aquifer portions, and the hydraulic response is measured in multi-level piezometers placed at different radial distances and in different directions to the pumping well. Extracting water from a partially penetrating well induces the vertical flow component required to resolve hydraulic anisotropy.

A synthetic study that resembles the following field tests was developed to find the best combination of model and data-acquisition strategy, while minimizing the uncertainty in the estimation of hydraulic anisotropy during model calibration. This study shows that there is a dependency between the model complexity and the required observation strategy which should be aligned with the modeling goals.

Field tests were performed in a fluvial gravel aquifer in the Upper Rhine Valley, Germany. In a series of pumping tests, water was sequentially extracted from three different intervals of a pumping well and the hydraulic response was observed at surrounding observation wells, placed at different distances and depths. A homogeneous anisotropic groundwater flow model and a heterogeneous one with multiple, locally anisotropic, horizontal layers, were fitted to the data collected during the field tests. Different measurement errors and model errors were considered by adapting the measurement uncertainties during model calibration. The results show that estimating differences in horizontal and vertical hydraulic conductivities is possible when data from pumping tests in a partially penetrating well and with a tomographic layout are inverted.

To better reproduce the true measurements and improve the estimation of hydraulic anisotropy, a multi-layer model that resolves the main vertical structure of hydraulic conductivity is preferable.

Zusammenfassung

Um Grundwasserströmungsprozesse verstehen zu können, werden detaillierte Informationen über die Heterogenität und Anisotropie der hydraulischen Durchlässigkeit benötigt. In Porengrundwasserleitern resultiert die räumliche Variabilität der hydraulischen Durchlässigkeit von der Schichtung unterschiedlichen Materials und besteht vorwiegend zwischen einzelnen horizontalen Lagen. Großräumig betrachtet induziert die vertikale Heterogenität hydraulische Anisotropie, für welche die horizontale Durchlässigkeit typischerweise größer ist als die vertikale Durchlässigkeit. Trotz der Bedeutung der hydraulischen Anisotropie, wie beispielsweise in der Auslegung hydraulischer Sanierungsmaßnahmen oder bei der Dimensionierung von Wasserhaltungsmaßnahmen in Baugruben oder Bergwerken, existieren bisher nur bedingt Forschungen zur Bestimmung des Unterschieds zwischen horizontaler und vertikaler Durchlässigkeit.

Diese Arbeit untersucht das Potential einer Feldmethode um die großräumige hydraulische Anisotropie induziert durch die vertikale Heterogenität in geschichteten Grundwasserleitern zu ermitteln. Der Ansatz basiert auf der Inversion von Absenkungsdaten die von Pumpversuchen in einem teilverfilterten Brunnen stammen, in welchem Grundwasser sequentiell aus verschiedenen Bereichen des Grundwasserleiters entnommen wird, und das hydraulische Antwortsignal in Mehrkanalmessstellen und Messstellenbündeln aufgezeichnet wird, die in verschiedenen Entfernungen und Richtungen zum Pumpbrunnen positioniert sind. Die Grundwasserentnahme aus einem teilverfilterten Brunnen induziert die vertikale Strömungskomponente die benötigt wird, um die hydraulische Anisotropie aufzulösen.

Es wurde eine synthetische Studie entwickelt, die den nachfolgenden Feldversuchen gleicht, um die beste Kombination von Modell und Datenerhebung zu bestimmen, bei gleichzeitiger Minimierung der Unsicherheit der ermittelten hydraulischen Anisotropie. Diese Studie zeigt, dass eine Abhängigkeit zwischen der Modellkomplexität und der benötigten Beobachtungsstrategie besteht, welche an die Modellziele ausgerichtet werden sollte.

Es wurden Feldversuche in einem fluviatilen Kiesgrundwasserleiter im Oberrheingraben, Deutschland, durchgeführt. In einer Reihe von Pumpversuchen wurde Grundwasser sequentiell aus drei verschiedenen Intervallen eines Pumpbrunnens entnommen und das hydraulische Antwortsignal wurde in verschiedenen Entfernungen und Tiefen zum Pumpbrunnen aufgezeichnet.

Es wurde ein homogen-anisotropes Grundwasserströmungsmodell und ein geschichtetes Modell mit mehreren, lokal-anisotropen, horizontalen Schichten verwendet, um die gemessenen Felddaten anzupassen.
Verschiedene Messfehler und Modellfehler wurden berücksichtigt indem die Messunsicherheiten während der Modellkalibrierung angepasst wurden. Die Ergebnisse zeigen, dass Unterschiede in den horizontalen und vertikalen Durchlässigkeiten ermittelt werden können, wenn Daten von Pumpversuchen in einem teilverfilterten Brunnen und mit einem tomographischen Layout invertiert werden. Um die gemessenen Daten besser wiedergeben zu können und die Schätzung der hydraulischen Anisotropie zu verbessern, empfiehlt es sich ein Mehrschichtmodell zu verwenden, welches die wesentliche vertikale Struktur der hydraulischen Durchlässigkeit auflöst.

Contents

1 Introduction **1**

1.1 Objectives 4

1.2 Thesis Structure 7

2 Theory **9**

2.1 Groundwater Flow Equation 9

2.2 Well Hydraulics 10

2.3 Anisotropy Ratio 11

2.4 Model Inversion 12

3 Joint Optimization of Measurement and Modeling Strategies **13**

3.1 General Approach 13

3.1.1 Definition of the Modeling Problem and Modeling Goals 15

3.1.2 Definition of Possible Observation Points 17

3.1.3 Definition of the Ensemble of Virtual Realities and of the Ensemble of Simplified Model Candidates 17

3.1.4 Mimicking a Best-Estimate Calibration 19

3.1.5 Optimization with Respect to the Modeling Goal 20

4 Synthetic Case Study **23**

4.1 Application to Radial Flow in Stratified Aquifers 23

4.1.1 Hydraulic Tests 24

4.1.2 Dewatering Scenario 24

4.1.3 Proposed Designs of Observation Wells in the Hydraulic Test 26

4.1.4 Generating Realizations of Hydraulic Conductivity as Virtual Realities and for Model Calibration 29

4.1.5 Numerical Implementation and Computational Effort 32

4.2 Results and Discussion 33

4.2.1 Reproduction of Measurements by the Calibrated Models 33

4.2.2 Predictive Performance of Calibrated Models 35
4.2.3 Predictive Errors 37
4.2.4 Multi-Objective Optimization 39

5 Field Application **43**
5.1 Field Site and Well Instrumentation 44
5.2 Hydraulic Tests 48
5.3 Data Processing 50
5.4 Processed Data Set 54
5.5 Numerical Model Setup 55
5.6 Model Calibration 58
5.7 Results and Discussion 61
5.7.1 Goodness of Model Calibration 61
5.7.2 Fitted Parameter Values 63
5.7.3 Assessing the Effective Conductivity Estimates 64
5.7.4 Comparison to Locally Isotropic Models 66
5.7.5 Updating the Model Setup 68
5.7.6 Considering Water Extraction from a Single Well Screen 71
5.7.7 Reducing the Number of Observation Points 76

6 Conclusions and Outlook **79**
6.1 Recommendations 83

List of Figures

3.1 General approach for jointly optimizing the measurement and modeling strategy of a complex system (from Maier et al., 2020). 15

4.1 (a) Setup of the hydraulic test to estimate hydraulic conductivity: Radial symmetric steady state groundwater flow to a partially penetrating well. Water is successively extracted along six well-screen sections i_{scr} at $r = r_w = 0.1$ m and drawdown is observed at all depths of water extraction in 3 out of 18 radial distances to the pumping well. Analysis of 816 designs, each design combining three multi-level piezometers. (b) Application to the dewatering of a construction pit: Radial symmetric steady-state groundwater flow (colored lines: head contours; black lines: streamlines) towards a construction pit within a large river. The base of the construction pit is at z_{pit}, sheet piles pushed to depth $z = z_{sheet}$ are placed at $r = r_{sheet}$ to separate the construction pit from the surrounding river. (from Maier et al., 2020) 25

4.2 Misplacing piezometers by perturbing the r- and z-coordinate of an intended measurement point in the virtual realities. The variability of the horizontal placement increases with depth. (from Maier et al., 2020) 27

4.3 Setup and sorting of designs by means of set 0 and set 1. In each set, the piezometer closest to the pumping well (green cross) is considered as fixed. The two remaining piezometers are shifted further outwards, starting with the outermost one (blue cross). After the outermost piezometer has reached the last radial distance available, the second piezometer (grey cross) is moved to the next position. This procedure is repeated until the second and third piezometer have reached the two last available positions. This is when the respective set is completed and the next set starts by moving the piezometer closest to the pumping well to the next position and so forth. (from Maier et al., 2020) . . . 28

4.4 Example of vertical profiles of horizontal and vertical hydraulic conductivity, K_r and K_z, respectively. (a) Highly resolved profile used as virtual reality; (b) partially upscaled profiles of K_r and K_z according to the models of 1, 2, 3, and 6 horizontal layers used as candidate fields in the calibration-mimicry. (from Maier et al., 2020) 31

4.5 Root mean square error $RMSE^y(d_j, k)$ of the best-fit of the 1-, 2-, 3-, and 6-layer models in meeting the drawdown differences of the virtual reality as function of the measurement design. Black line: root mean squared values of the measurements. Purple line: Measurement error expected from the measurement devices employed in field tests. (from Maier et al., 2020) 34

4.6 Fraction of all realizations n_r in which a particular model candidate was determined as best among the available model candidates in minimizing $E^t_{rel}(d_j, k, l)^2$ of the target quantities: (a) Q_d, (b) K_r^{eff} and (c) K_z^{eff}. (from Maier et al., 2020) 36

4.7 Root mean squared relative errors $RMSE^t_{rel}$ in predicting K_r^{eff} (blue), K_z^{eff} (red), and Q_d (black) as function of the chosen measurement designs for individual model choices. (a) 1-layer model, (b) 2-layer model, (c) 3-layer model, (d) 6-layer model. (from Maier et al., 2020) 38

4.8 Optimized modeling and measurement strategies for reducing the relative prediction error of target quantities. The color of the points refers to the resulting model candidate, i.e., gold for the 6-layer model and blue for the 3-layer model. a. Results for minimizing errors of each target quantity K_r^{eff}, K_z^{eff}, and Q_d, respectively. Choosing one of these designs and models implies making a compromise with respect to minimizing the prediction uncertainty of the respectively two other target quantities. b. Designs resulting from multi-objective optimization. (from Maier et al., 2020) 40

4.9 Root mean squared relative errors $RMSE^t_{rel}$ in predicting the three target quantities for all measurement designs and all model choices (indicated by colors). (a) pairwise plot of $RMSE^t_{rel}$ for $t = K_r^{eff}$ and $t = K_z^{eff}$, (b) pairwise for $t = K_r^{eff}$ and $t = Q_d$, (c) pairwise for $t = K_z^{eff}$ and $t = Q_d$, (d) $RMSE^t_{rel}$-values for all three targets. Gray lines in (a)-(c) and planes in (d): separation of the 10% best from the rest. (from Maier et al., 2020) 41

5.1 a. Overview map of the field site. b. Plan view of the installed measurement network at the site. 45

5.2 a. Side view and b. plan view of the grab drilling of pumping well R01. 46

5.3 Overview of the measurement network installed at the field site. The vertical coordinate is measured from the base of the aquifer, the horizontal coordinates from the pumping well R01 in the middle, illustrated by its well screens. 47

5.4 Estimation of the well efficiency of pumping well R01 (Stoll, 2020). 48

5.5 Configurations of the straddle packer according to water extraction from well-screen section I, II or III. 49

5.6 Exemplary datasets excluded from model calibration due to instable pressure transducer recordings. 51

5.7 Dataset of pumping test $p_t = 5$ when water was extracted from the upper screen (hydraulic test I) showing a. the absolute drawdown measurements versus pumping time. b. the difference in drawdown measured in the different observation wells and reference well W04.3. c. the logarithmic time derivative of drawdown. 53

5.8 Field measurements of the hydraulic tests performed at (a) the top, (b) the middle and (c) the bottom screen of the pumping well. The yellow, grey, blue and orange colors correspond to the radial distances $r_1 \in$[3.3,3.6], $r_2 \in$[6.2,6.8], $r_3 \in$[10.2,10.7] and $r_4 \in$[20.9,21], respectively. 55

5.9 a. Measurement errors σ_{repr} resulting from the reproducibility analysis as function of the drawdown measurements s_{meas}. Colors indicate extraction from the different pumping intervals in different depths, marker styles indicate horizontal orientation of the measurement points. b. Measurement errors σ_{c_p} resulting from measurements at similar radial distances and depths but different directions to the pumping well. 56

5.10 Conceptual representation of the 1-layer and 5-layer model, distinguished by dashed horizontal lines. The drilling profile is displayed left to the pumping well showing lithologic information at the pumping well location. Also, the gravel pack (red) and clay seal (gray) are displayed right to the pumping well. The black crosses correspond to individual observation points, placed at radial distances $r_1 \approx 3.5$ m, $r_2 \approx 6.5$ m, $r_3 \approx 10.5$ m and $r_4 \approx 21$ m. The model domain reaches from the saturated aquifer thickness at $z_{top} = 41.05$ m to the aquifer bottom at $z_{bot} = 0$ m. Note that axes are not to scale and that the contour lines of drawdown correspond to an exemplary response in a homogeneous aquifer system, i.e. considering a 1-layer model. 57

5.11 a. Relative gravel and sand fractions from the field description of the drilling core of the large diameter well R01 at the field site Kappel-Grafenhausen (white gaps = core loss). b. Definition of five horizontal layers considering depths of more than half a meter in thickness in which the sand fraction clearly exceeds the gravel fraction. c. Calibrated radial (blue) and vertical hydraulic conductivities (red) of the 1-layer (dashed lines) and 5-layer model (solid lines). . . . 58

5.12 Assessment of model results: Absolute difference between simulated and measured drawdown versus the measured drawdowns of the 1-layer and 5-layer model and the thereto fitted error models (a and b). Field measurements versus simulated results of the 1- and 5-layer models (c and d) with errorbars according to the error model. The black diagonal lines represent the 1:1 identity lines. 62

5.13 Simulated drawdowns versus measured drawdowns when running the 1-layer model with the effective conductivity estimates obtained from the 5-layer model calibration. The black diagonal line present the 1:1 identity line. 65

5.14 Assessment of model results when considering isotropic hydraulic conductivity: Absolute difference between simulated and measured drawdown versus the measured drawdowns of the 1-layer and 5-layer model and the thereto fitted error models (a and b). Field measurements versus simulated results of the 1- and 5-layer models (c and d) with errorbars according to the error model. The black diagonal lines represent the 1:1 identity lines. 67

5.15 Absolute difference between simulated and measured drawdown versus the measured drawdown of the 3-layer model and the thereto fitted error model (a). Field measurements versus simulated results of the 3-layer model (b) with errorbars according to the error model. The black diagonal line represents the 1:1 identity line. 69

5.16 Field measurements versus simulated results of the 1-layer model when separately inverting data from hydraulic test I (a), II (b), and III (c) with errorbars according to the error model. Field measurements versus simulated results of the 3-layer model when separately inverting data from hydraulic test I (d), II (e), and III (f) with errorbars according to the error model. The black diagonal lines represent the 1:1 identity lines. 72

5.17 Radial and vertical conductivities K_r and K_z estimated for each available horizontal layer of the 1-layer model (a-b) and 3-layer model (e-f) considering the following inversion approaches: inverting data from pumping tests following the tomographic approach, inverting data from hydraulic test I (water extraction from the upper screen), II (water extraction from the middle screen), and III (water extraction from the lower screen), and inverting a reduced data set. Associated standard deviations of log-conductivities resulting from the 1-layer (c-d) and 3-layer model calibration (g-h) using different inversion approaches. 74

5.18 Field measurements versus simulated results of the 1-layer model (a) and 3-layer model (b) considering data inversion of tomographic pumping tests with a reduced dataset and with errorbars according to the error model. The black diagonal lines represent the 1:1 identity lines. 77

List of Tables

1.1 Proportions of collaborative work presented in Chapter 3 and 4. 7

4.1 Parameters of the tests case. 30

5.1 Key data of the pumping test series. 49

5.2 Details on the sensors used to measure the drawdown response to water extraction in pumping well R01. 50

5.3 Calibrated radial and vertical hydraulic conductivities and the associated standard deviations $\hat{\sigma}$ of estimation of each horizontal layer in the 1-layer and 5-layer models. 64

5.4 Calibration results of the locally anisotropic 1-layer and 5-layer models. 64

5.5 Calibrated isotropic hydraulic conductivities and the associated standard deviations $\hat{\sigma}_{lnK}$ of estimation of each horizontal layer in the 1-layer and 5-layer models. Error model parameters of the isotropic 1-layer and isotropic 5-layer model. 68

5.6 Calibrated radial and vertical hydraulic conductivities and the associated standard deviations $\hat{\sigma}$ of estimation of each horizontal layer in the 3-layer model. . 70

5.7 Calibration results of the locally anisotropic 3-layer model. 71

5.8 Calibration results of the locally anisotropic 1-layer model when separately inverting data of hydraulic test I, II, and III. 75

5.9 Calibration results of the locally anisotropic 3-layer model when separately inverting data of hydraulic test I, II, and III. 75

5.10 Calibration results of the locally anisotropic 1-layer and 3-layer model considering a reduced dataset. 78

Chapter 1

Introduction

Groundwater flow and transport processes in porous media are part of a broad research field studied by hydrogeologists. Characterizing these processes requires abundant information on the subsurface features and their hydraulic properties, with hydraulic conductivity **K** being one of the most important. Hydraulic conductivity quantifies the ease with which groundwater flows through a porous medium, and depends on both the properties of water, and the physical properties of the aquifer material, such as the pore size, geometry and connectivity (Freeze and Cherry, 1979). In natural systems, these properties change in space, leading to spatial variations of hydraulic conductivity (heterogeneity). The hydraulic conductivity also depends on the direction of flow (anisotropy), which is strongly influenced by the sediment bedding that result from variable sediment transport and deposition regimes (Bennett et al., 2019). The study of hydraulic anisotropy is a cornerstone of this work. The effects of hydraulic anisotropy on groundwater flow are often misconceived and confused with that of heterogeneity. To clearly distinguish both concepts, they are defined in the following.

Subsurface heterogeneity is the spatial variability of subsurface properties as the result of non-uniform and dynamic conditions under which material deposition occurs. Subsurface material may express individual, continuous or intermittent layers, or include lenses composed of material that differ from the surrounding (Kruseman and de Ridder, 1994). In sedimentary aquifers, the subsurface structure is the imprint of depositional processes occurring at different velocities and geologic timescales that determine the size and amount of sediment load (Aigner et al., 1998; Bennett et al., 2017). As a

result, the subsurface is of bedded character in which hydraulic conductivity especially varies among individual strata, leading to vertical heterogeneity.

In contrast to heterogeneity, anisotropy refers to the directional dependence of material properties, and its expressions are closely related to the observed scale. On small scales, the directional dependence on hydraulic conductivity is a result of the orientation of grains, which are typically elongated in the horizontal direction (Borghi et al., 2015). On larger scales, the continuity of channel deposits in e.g., alluvial aquifers with internal variation of their hydraulic properties, can lead to strong horizontal anisotropy (Gianni et al., 2019). In stratified aquifers, the bedded character enhances groundwater flow along layer boundaries, rather than across them (Bear, 1972; Borghi et al., 2015), inducing a higher horizontal than vertical hydraulic conductivity. Such aquifers can be described by an effective hydraulic conductivity tensor $\mathbf{K_{eff}}$ with the principal components aligned with the horizontal and vertical directions. The anisotropy ratio, i.e. the ratio of horizontal to vertical hydraulic conductivity, is typically larger than unity, unless the layers are tilted (Bear, 1972; Maier et al., 2020), and can reach ratios of up to 100 in stratified aquifers (Freeze and Cherry, 1979).

Many hydrogeological applications, such as the design of remediation systems (Cardiff and Barrash, 2011; Zschornack et al., 2013; Bair and Lahm, 1996; Zlotnik and Ledder, 1996), the prediction of solute transport (Bohling et al., 2007; Sanchez-Leon et al., 2016), and the enhancement of transverse mixing of solute plumes (Chiogna et al., 2015; Cirpka et al., 2015), depend not only on precise information on subsurface heterogeneity but also require information on hydraulic anisotropy. Particularly relevant activities in which information on the hydraulic anisotropy is essential, are the delineation of capture zones of partially penetrating wells (Bair and Lahm, 1996), the design of horizontal collector wells for water supply, and the application of dewatering measures near big rivers to keep a construction pit dry (Maier et al., 2020). Hydraulic anisotropy influences the spatial extent of the depression cone when lowering water tables upon pumping. While higher hydraulic conductivities in the horizontal direction increase the lateral extent of the depression cone, the comparatively lower vertical hydraulic conductivities limit its vertical extent (Maier et al., 2020; Bair and Lahm, 1996). Neglecting the ratio of horizontal to vertical hydraulic conductivity in the calculation of such systems can lead to an inefficient design and unnecessary costs.

Several experimental methods have been developed for resolving the spatial variability of hydraulic conductivity at different scales and degrees of resolution. Direct push methods, such as direct-push injection logging (DPIL; Butler et al., 2002; Dietrich et al., 2008; Lessoff et al., 2010) and the direct–push permeameter (DPP; Butler et al., 2007; Chen et al., 2008, 2010; Klammler et al., 2011; Zschornack et al., 2013) can resolve the variation of hydraulic conductivity in the vertical direction. In these tests, a steel rod is advanced vertically into the subsurface with a short filter section attached to its lower end. At depths in which information on hydraulic conductivity is desired, the tool is halted and measurements are performed based on water injection to the surrounding aquifer portion. While direct-push injection logging is based on measuring the hydraulic resistance of the aquifer material surrounding the screen attached to the steel rod by considering the ratio between the pressure applied and the discharge (Dietrich et al., 2008), the direct-push permeameter measures the discharge and head response at two locations above the injection screen (Butler et al., 2007). Both methods can provide one-dimensional vertical hydraulic conductivity profiles at high resolution.

Another well established field method is hydraulic tomography, which aims to resolve subsurface heterogeneity at large scales and high resolutions (Gottlieb and Dietrich, 1995; Yeh and Liu, 2000; Bohling, 2009; Cardiff et al., 2009). Hydraulic tomography consists of a series of pumping tests in which different aquifer portions are stressed either by injection or extraction of water, and the hydraulic head responses are recorded at numerous observation points. Despite the abundant information on the vertical variation of hydraulic conductivity contained in data collected by hydraulic tomography, the reported applications concentrate on the characterization of aquifer heterogeneity rather than focusing on anisotropy.

Despite the importance of hydraulic anisotropy, only few studies present a rigorous assessment of the ratio of horizontal to vertical hydraulic conductivity on large scales. Klammler et al. (2017) proposed a theoretical framework to estimate the bulk hydraulic anisotropy from measurements obtained with the direct-push permeameter, however, it does not consider the vertical variability of hydraulic conductivity.

An alternative method is the tomographic slug test proposed and tested by Paradis et al. (2015, 2016) in a littoral aquifer. Paradis et al. (2015, 2016) suggest that tomographic slug tests are more appropriate for resolving hydraulic anisotropy induced by heterogeneities at small scales, also in the horizontal direction, e.g., from cross-bedding.

Similar to hydraulic tomography, in tomographic slug tests different aquifer portions are stressed in a series of slug tests, while the hydraulic head responses are measured in surrounding observation points. A limitation of this field method is the very small range of investigation in the horizontal direction (typically <10m), especially in highly permeable aquifers (Paradis et al. 2015).

Overall, there is still a lack of research investigating the distinction between horizontal and vertical hydraulic conductivities in highly permeable and heterogeneous aquifers, and their effects on groundwater flow.

1.1 Objectives

In this work, I focus on the concept of hydraulic anisotropy on large scales and its interrelation with the spatial variability of hydraulic conductivity in stratified aquifers. Among the previously reviewed field methods, hydraulic tomography presents a very promising approach for resolving vertical variations of hydraulic conductivity. Considering the informative value of tomographic data with respect to scale and resolution, I raise the question to what extent pumping tests with a tomographic setup provide a potential method for determining the directional dependence of hydraulic conductivity on large scales.

The main objective of this thesis therefore is to investigate the potential of a hydraulic tomographic approach to estimate the large-scale hydraulic anisotropy induced by the vertical heterogeneity on smaller scales in stratified aquifers. In the following, I define guiding research questions, hypothesis, and methodologies to address the overall goal of this thesis and assign them to four different work packages.

Work Package 1: Joint Optimization of Measurement and Modeling Strategies

In the first work package, I address the question of how to setup and select a numerical groundwater flow model as an appropriate representation of the true system. At the same time, I want to encounter an optimal data-acquisition strategy to inform the model with non-redundant and conclusive data when bearing an available time and budget in mind. With this, the main research question arises of whether the data should be collected based on a predefined model or vice versa.

Typically, both model selection and data collection aim to reduce the prediction uncertainty of unknown parameters but concern two different disciplines in hydrogeology. From a modelers point of view, the main hypothesis would state that model performance can be improved by finding the optimal model. Field operators on the other hand would hypothesize that feeding a model with the right data is key for estimating parameters with small uncertainty. Usually, model selection and data collection are considered individually, in many cases simply because either the model or the experimental setup and/or field data are already available. Unfortunately, this may lead to the implementation of sub-optimal models and/or field tests. I therefore hypothesize that finding the best model and best data-acquisition strategy is a joint problem calling for a joint solution. I propose a joint optimization that is able to find the optimal combination of model and data-acquisition strategy to improve model simulations of a natural system. Though it can be applied to various modeling problems, I restrict its application to improve the estimation of hydraulic anisotropy. For a proof-of-concept, I analyze hypothetical pumping tests with a tomographic setup. The proposed method is published by Maier et al (2020).

Work Package 2: Establish a Research Field Site

The second work package deals with the implementation of pumping tests that follow the principles of hydraulic tomography. To test a field application of the tomographic-pumping tests, similar to the synthetic field tests in work package 1, a new research site was established in Kappel-Grafenhausen, located in the Upper Rhine Valley, Germany. The goal of this work package is to design a well network at the field site that facilitates the estimation of anisotropic hydraulic conductivity. For example, decisions must be made regarding the design of the pumping well and the number and placement of observation wells.

The site was instrumented with a large diameter pumping well that was equipped with three isolated well-screen sections to facilitate water extraction from different depths. To measure depth-oriented hydraulic responses to pumping, the observation wells were installed as clusters of partially penetrating wells and multi-level wells. The well clusters and multi-level wells were placed at different radial distances and directions to the pumping well.

I performed a total of 22 pumping-test series in which water was successively extracted

from the three well-screen sections of the partially penetrating pumping well and a total number of 58 observation points were monitored during each test.

Work Package 3: Develop a data-processing strategy to generate a comprehensive data base from hydraulic tests

In total, more than a thousand drawdown curves were obtained from the performed pumping tests. This raises the question of how to handle large data amounts obtained from field applications and how to condense the data for use in posterior data inversion. In general, I hypothesize that the principle of data reproducibility is a useful concept for data validation and data reduction. Another hypothesis is that the challenges related to analyzing the complete transient response of the data or of reaching a steady-state pumping regime can be circumvented by considering the steady-shape regime of the pumping tests.

To facilitate the use of the large amount of data, I developed a data-processing strategy that reduces the data set to a manageable size while clearly exposing the reproducibility of the pumping-test data.

Work Package 4: Assessment of the model performance

Inverting the data collected at the field site requires a groundwater flow model with a suitable representation of the aquifer under investigation. In the fourth work package, I address the research question of how to setup a model that ideally has the smallest possible level of complexity to represent the true system and how to evaluate model performance. Along these lines, I hypothesize that it is indispensable to use existing hydrogeological information of a field site for adapting the model setup to a specific site. As different models have different errors, I hypothesize that model performance should be assessed by considering not only different measurement errors but also the model error.

For inverting the pumping test data, I tested the performance of two different models with varying degrees of complexity. A simple groundwater flow model was defined as homogeneous but anisotropic, whereas a more complex variant included five individual and locally anisotropic horizontal layers. The layering was defined based on lithological information obtained from the drilling profile at the pumping-well location. To account for random, systematic, and conceptual model errors, I propose a calibration strategy where the total error is determined during the calibration process. With this, the model performance is assessed based on the model error required to accept a model.

1.2 Thesis Structure

In the following Chapter 2, I present the underlying theory including the governing equations for groundwater flow, well hydraulics and the concept of anisotropy considered in this work. In Chapter 3, I present the methodology to jointly optimize the measurement and modeling strategies and apply it to a synthetic test case considering radial flow in stratified aquifers for predicting the large-scale hydraulic anisotropy in Chapter 4. Chapter 5 describes the field site that was instrumented in the course of this work, with details on the hydrogeological setting and the measurement network. Chapter 5 also describes the design of the pumping tests performed at the field site, presents an overview of the database generated with the processed dataset, the methodology followed to assess data reproducibility, and the error model adopted in this work. The final part of Chapter 5 presents the model inversion with the field data including a description of the numerical model and the calibration strategy, and a discussion of the results. In Chapter 6, I summarize the motivation and findings of my work and draw conclusions on the potential and limitations of tomographic pumping tests for resolving hydraulic anisotropy on large scales. Additionally, I give recommendations to improve the application of the field tests for enhancing the estimation of anistropic conductivity.

The work presented in Chapter 3 and 4 has essentially been published in *Water Resources Research* (Maier et al., 2020). Table 1.1 lists the proportion of contributions associated with each author of the publication.

Table 1.1: Proportions of collaborative work presented in Chapter 3 and 4.

Author	Author position	Scientic Ideas %	Data Generation %	Analysis & Interpretation %	Paper Writing %
Ruth Maier	First	20	95	35	30
Ana Gonzalez-Nicolas	Second	10	0	5	10
Carsten Leven	Third	20	0	10	10
Wolfgang Nowak	Fourth	25	0	20	20
Olaf A. Cirpka	Fifth	25	5	30	30
Title of paper:	Joint Optimization of Measurement and Modeling Strategies With Application to Radial Flow in Stratified Aquifers				
Publication Status:	Published				

Chapter 2

Theory

2.1 Groundwater Flow Equation

Following the law of mass conservation, the total source and sink of mass in a control volume equals the sum of the change in mass storage and the net mass flux across the surface of the control volume. In a saturated porous medium, the entire pore space is saturated with water and the effective porosity n_e equals the volumetric water content. Combining Darcy's law and the continuity equation yields the following governing equation for transient groundwater flow in a porous medium:

$$S_0 \frac{\partial h}{\partial t} - \nabla \cdot (\mathbf{K} \nabla h) = W_0 \tag{2.1}$$

where W_0 denotes the total sources and sinks, h denotes the hydraulic head, t is time and ∇h describes the hydraulic gradient. The specific storage coefficient S_0 is equal to the sum of the compressibility of the pore space and water:

$$S_0 = \frac{\partial n_e}{\partial h} + \frac{n_e}{\rho_w} \frac{\partial \rho_w}{\partial h} \tag{2.2}$$

where ρ_w denotes the density of water.

In eq. (2.1) $\mathbf{K}$ is the hydraulic conductivity tensor

$$\mathbf{K} = \begin{bmatrix} K_{xx} & K_{xy} & K_{xz} \\ K_{yx} & K_{yy} & K_{yz} \\ K_{zx} & K_{zy} & K_{zz} \end{bmatrix} \tag{2.3}$$

where the off-diagonal entries become zero if the principle directions are aligned with the x-, y- and z-directions of the applied coordinate system.

2.2 Well Hydraulics

Radial-symmetric groundwater flow to a partially penetrating pumping well can be expressed in radial and vertical coordinates r and z, respectively:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(K_r r \frac{\partial h}{\partial r}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial h}{\partial z}\right) = S_0 \frac{\partial h}{\partial t} \tag{2.4}$$

where K_r and K_z denote the radial and vertical hydraulic conductivities and h is the hydraulic head. By assuming a horizontal layering in stratified aquifers, I consider the radial and vertical hydraulic conductivities of the aquifer to vary only in the vertical direction, i.e. $K_r(z)$ and $K_z(z)$.

In pumping-test analysis usually the drawdown s is considered which is the change in hydraulic heads induced upon pumping. By measuring the drawdown rather than the hydraulic head, the influence of ambient flow is eliminated. Considering the drawdown, eq. (2.4) can be written as

$$\frac{1}{r}\frac{\partial}{\partial r}\left(K_r r \frac{\partial s}{\partial r}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial s}{\partial z}\right) = S_0 \frac{\partial s}{\partial t} \tag{2.5}$$

In all numerical simulations, I consider confined conditions and a flat aquifer base and top. The upper and lower boundaries z_{top} and z_{bot} are no-flow boundaries

$$\left.\frac{\partial s}{\partial z}\right|_{z=z_{bot}} = \left.\frac{\partial s}{\partial z}\right|_{z=z_{top}} = 0 \,\forall r \tag{2.6}$$

and zero drawdown occurs at all depths at the outer radius R of the domain

$$s(R, z) = 0 \, \forall z \tag{2.7}$$

In all hydraulic tests presented in this work, water is extracted from different depths of a partially penetrating well. The boundary conditions along the well with radius $r = r_w$ are defined as the following

$$\begin{cases} s = \text{constant and} & \\ -\int_{z_c(i_{scr})-\frac{w}{2}}^{z_c(i_{scr})+\frac{w}{2}} 2\pi r K_r \dfrac{\partial s}{\partial r} dz = Q_w & \text{if} \quad z_c(i_{scr}) - \frac{w}{2} \leq z \leq z_c(i_{scr}) + \frac{w}{2} \\ \dfrac{\partial s}{\partial r} dz = 0 & \text{otherwise} \end{cases} \tag{2.8}$$

assuming a constant head and a constant total flux Q_w along the well screen i_{scr} considered for water extraction. No flow boundaries are assigned at all depths below the active well screen of length w, centered at z_c.

Instead of considering the transient regime of the pumping tests, which would require specifying initial conditions and the specific storativity, I analyze the constant-shape regime of the pumping tests, in which the absolute values of drawdown still change but the hydraulic-head differences between measurement locations remain constant (Bohling et al., 2002, 2007). The drawdown differences between observation points can therefore be simulated by the steady-state drawdown equation:

$$\frac{1}{r}\frac{\partial}{\partial r}(K_r r \frac{\partial s}{\partial r}) + \frac{\partial}{\partial z}(K_z \frac{\partial s}{\partial z}) = 0 \tag{2.9}$$

2.3 Anisotropy Ratio

The **K**-field in stratified aquifers can be described by highly-resolved vertical conductivity profiles in which the radial and vertical hydraulic conductivities K_r and K_z vary among individual horizontal layers. I upscale the horizontal conductivity by taking the arithmetic average of the highly resolved K_r-field over the layer thickness, and the

vertical conductivity by the harmonic average of the highly-resolved K_z-field:

$$K_r^{eff}(j) = \frac{1}{\Delta z} \int_{z_{bot}(j)}^{z_{top}(j)} K_r(\zeta) d\zeta \tag{2.10}$$

$$K_z^{eff}(j) = \Delta z \int_{z_{bot}(j)}^{z_{top}(j)} \left(\frac{1}{K_z(\zeta)} d\zeta \right)^{-1} \tag{2.11}$$

in which $z_{bot}(j)$ and $z_{top}(j)$ are the bottom and top vertical coordinates of the layer j over which the averaging is performed, respectively, and $\Delta z = z_{top}(j) - z_{bot}(j)$ is the corresponding layer thickness.

The ratio of K_r^{eff} to K_z^{eff} is the anisotropy ratio

$$\vartheta = \frac{K_r^{eff}}{K_z^{eff}} \tag{2.12}$$

2.4 Model Inversion

Knowledge of the hydraulic properties of a system is often scarce when modeling groundwater flow. A common method to estimate unknown parameters is the approach of numerical inverse modeling in which the unknown parameters $\mathbf{m}$ of a forward model are estimated by matching the dependent variables $\mathbf{y}$ to field observations $\mathbf{d}$ (Poeter and Hill, 1997).
In this thesis, I determine the unknown model parameters by performing a non-linear least squares fit and considering an error model that depends on the model outcome. In general, the objective function states

$$\phi = (\mathbf{d} - \mathbf{y}(\mathbf{m}))^T \mathbf{C}_{\mathbf{yy}}^{-1} (\mathbf{d} - \mathbf{y}(\mathbf{m})) \tag{2.13}$$

with $\mathbf{C_{yy}}$ being the covariance matrix of errors. The model inversion then implies finding the parameter set $\mathbf{m}$ that minimizes the objective function ϕ. This requires that the number of independent measurements, i.e. the length of the data vector $\mathbf{d}$, is larger than the number of parameters, i.e. the length of the parameter vector $\mathbf{m}$.

Chapter 3

Joint Optimization of Measurement and Modeling Strategies [1]

3.1 General Approach

The process of scientific modeling for environmental applications has been widely discussed by the scientific community, and big efforts have led to a large number of publications that focus on its formalization (e.g., Hill and Tiedeman, 2007; Walker et al., 2003; Oreskes, 1998; Woessner and Anderson, 1996). Though differing in some details, all these formalizations agree that model construction and data acquisition are intertwined and that the typical modeling cycle consists of clearly defining the modeling goal, constructing a conceptual model, translating it into a numerical model, collecting informative data, calibrating and validating the model, and if several competing models have been formulated, selecting the best one of them.

There are two main challenges in this cycle: The first one is to decide on the most appropriate modeling strategy: where a balance between model complexity and reasonable simplifications should be pursued. This is a recurring question of model selection (Beven, 2002; Raftery, 1995), that also rises frequently in groundwater problems (e.g., Yeh and Yoon, 1981). The second challenge is to decide on appropriate strategies to collect data. How much and which data is sufficient and conclusive (for calibration, validation, and selection) while keeping the costs for experimental/field campaigns at a

[1] The essential content of Chapter 3 is published by Maier et al. (2020)

bearable level? This well-known problem is called optimal design of experiments. It is mostly formalized as a mathematical optimization problem under the uncertainty of the yet unmeasured data (e.g., Leube et al., 2012; Müller, 2005; Chaloner and Verdinelli, 1995).

Traditionally, these two challenges are treated as disjoint: either one selects a model with data being given, or one optimizes a field campaign to feed the information needs of a given model. Therefore, an even more substantial question regarding the modeling cycle arises: should one design the measurement strategy based on the selected model or vice versa? Apparently, there is an interrelation of measurement and modeling strategies, which rather calls for a joint solution. So far a rational, objective approach that jointly assists both modelers and experimentalists to target the joint problem more systematically has been lacking.

I propose that rational model selection should go hand-in-hand with cost-efficient data collection to achieve the best possible model performance towards a given modeling goal for a given experimental budget. I hypothesize that this joint problem can be solved by a corresponding joint optimization. This optimization has to find the best combination of model choice and corresponding field-campaign design, so that the best expected performance towards the given modeling goal can be achieved. As predictive model performance is only speculative during the design phase of models and field campaigns, the optimization needs a means of anticipating both predictive uncertainty and systematic predictive errors. To include both uncertainty and systematic errors in this step, I propose to use an ensemble of highly resolved virtual realities to generate realistic virtual measurements and prediction targets. While these virtual realities provide hypothetical truths, calibrating models with this level of detail would trigger infeasible data requirements. Along with the virtual realities, ensembles of simplified models with different model complexity are generated that undergo model selection for different measurement designs that are subject to choice.

In the following, I present the framework for the joint optimization. The methodology includes five steps which are illustrated in Figure 3.1.

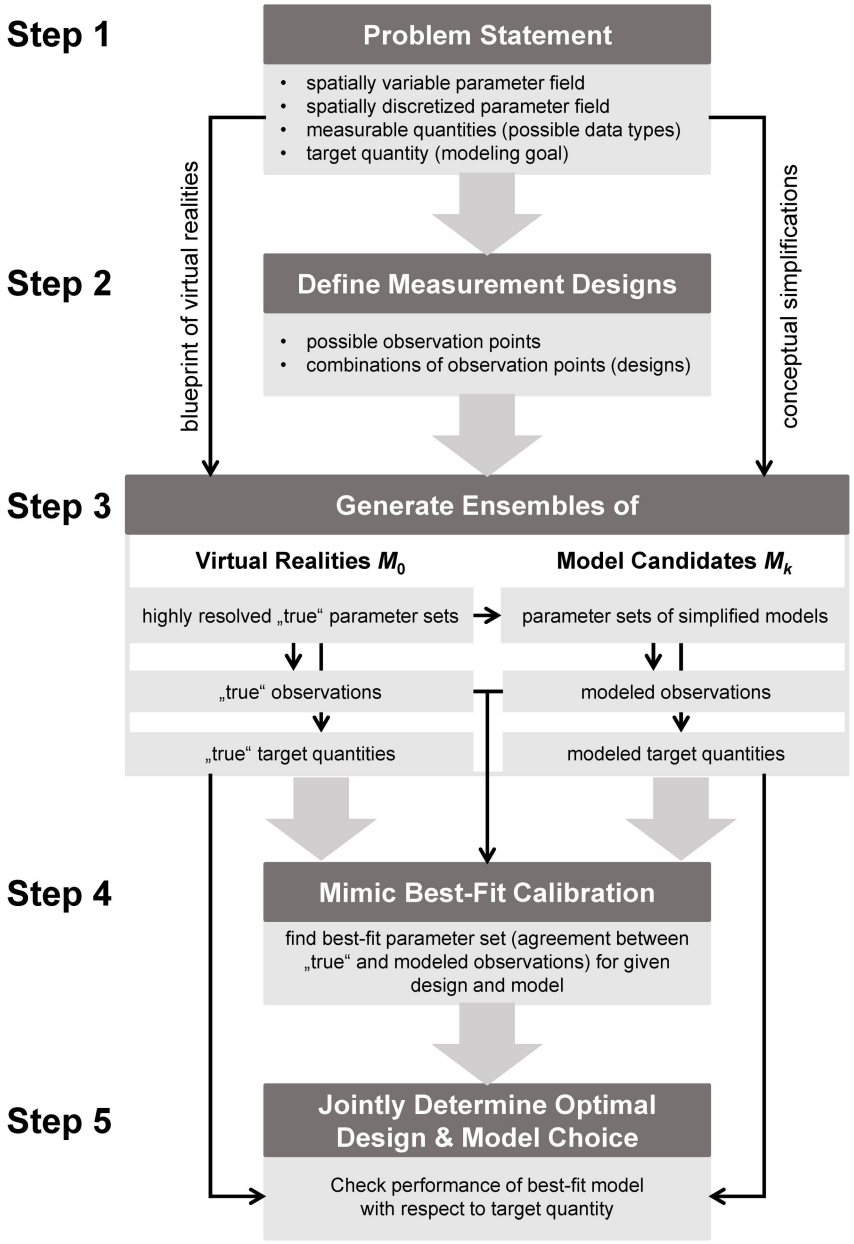

Figure 3.1: General approach for jointly optimizing the measurement and modeling strategy of a complex system (from Maier et al., 2020).

3.1.1 Definition of the Modeling Problem and Modeling Goals

The first step includes defining the basic conceptual model of a problem which involves stating the governing equations with their initial and boundary conditions to simulate

the system at hand. Conceptual uncertainties may arise from the selection of state variables and processes to be modeled, from the spatial discretization of the material properties of the system and from the initial and boundary conditions of the modeling problem (Tiedeman et al., 2003). For example, in this thesis the conceptual uncertainty is restricted to the scale and zonation of material properties only.

The true spatially variable parameter field is denoted x, which in reality exhibits variations on all spatial scales. In order to achieve identifiability in model calibration, one must typically discretize or simplify the parameter field, for example by the introduction of internally homogeneous zones, or the definition of a limited set of spatial base functions used for spatial interpolation (e.g., RamaRao et al., 1995). This yields a vector $\mathbf{x}_k$ of a few uncertain parameter values, in which the index k refers to a model choice (see Section 3.1.3). Hence, both the uncertainty of parameter values and the structural uncertainty are considered regarding the spatial discretization of the parameter fields. These two uncertainties are not independent of each other – it is well known that a coarser representation of material properties requires partially upscaled (effective) parameter values (Rubin, 2003, Chapter 5), if not even effective governing equations that differ from the equations used at high resolution.

In order to calibrate the competing models, a quantity y must be defined which could potentially be measured in a field test and can be computed by the model candidates. Choosing a measurement design, that is, a set of locations where to measure y, will be discussed in Step 2 (see Section 3.1.2).

The performance of the calibrated models, for given measurement designs and parameter representation, requires defining a quantitative and computable target quantity t associated with a given modeling goal. An optimized measurement and modeling strategy should lead to the best agreement in the target quantity between the prediction of the calibrated model and observations of the target quantity. At the moment when the model and experimental strategy are developed, however, true measurements do not exist yet, so that they have to be simulated with the given uncertain knowledge about the system.

The last aspect is how to handle the high prior uncertainty regarding the true system. The identified best measurement setup and conceptual model should be robust against this uncertainty. Therefore, an ensemble of virtual realities at high spatial resolution is considered and the performance of the measurement designs and model concepts is

computed over the entire ensemble. This ensemble of virtual realities allows mimicking many possible outcomes of any proposed measurement design, and so to statistically anticipate the subsequent calibration, validation and model selection.

3.1.2 Definition of Possible Observation Points

In Step 2 of the proposed framework, a set of possible observation locations, here called points, needs to be defined. By this, the optimal-design problem becomes a selection process among preset designs. The individual possible points are denoted $p_i, i = 1 \ldots n_p$, where n_p is the total number of observation points considered. This set is larger than the number of observation points that could be assessed in a single, feasible field experiment. Thus, many preset designs can be proposed as different subsets from this large set: permissible measurement combinations are defined, denoted as the designs $d_j, j = 1 \ldots n_d$ to be tested, in which j is an index to a particular design, and n_d is the total number of designs. These combinations are chosen to be realistic from a field-experimental perspective regarding the number of points and their spatial arrangement. Note that the number of observation points chosen in design d_j, denoted n_{obs}, can be design specific.

In classical optimal design, the problem would be to identify the best measurement design d_j within the set of proposed designs, given a conceptual model M_k (see Section 3.1.3), minimizing an uncertainty metric of the target quantity t. Instead, the error metric is minimized between a yet-to-be selected model M_k and an assumed (yet uncertain) truth.

3.1.3 Definition of the Ensemble of Virtual Realities and of the Ensemble of Simplified Model Candidates

Step 3 includes the definition of several model ensembles: The ensemble of virtual realities M_o replaces the true field system. These models exhibit the highest possible resolution of the parameter field x and include all known processes. The discretized parameter field of the virtual reality is denoted as $\mathbf{x}_0$. Because the prior uncertainty is fairly high, many virtual realities are generated, leading to an ensemble of $\mathbf{x}_0$. In

the following, $\mathbf{x}_0(l), l = 1 \ldots n_r$, is the parameter set of realization l associated with the virtual reality $M_o(l)$, and n_r is the number of realizations.

In addition to the virtual realities M_o, n_m ensembles of feasible model candidates $M_k, k = 1 \ldots n_m$ are specified. These are conceptually simpler than the virtual realities and could actually be calibrated in real applications (Doherty and Christensen, 2011). In the application presented in Chapter 4, the individual ensembles M_k differ from each other by the resolution (and hence scale) of the parameter fields x.

In principle, generating the n_m ensembles of models M_k could be independent of the virtual realities, but in this work they are constructed by upscaling the parameter vectors $\mathbf{x}_0(l)$ of each virtual reality l to a lower resolution, resulting in one upscaled parameter vector $\mathbf{x}_k(l)$ for each simplified model k and realization l. This procedure ensures that the different model ensembles are realistically similar, and it avoids the difficult procedure of specifying consistent parameter priors for models at different scales. Using all $\mathbf{x}_0(l)$ and $\mathbf{x}_k(l)$, hypothetical field data (measurable quantities y) are simulated at all possible observation locations p_i considered. In the following, $y_0(p_i(d_j), l)$ is the (error-free) simulated measurement at point i within design j in virtual reality (or realization) number l, whereas $y_k(p_i(d_j), l)$ is the same simulated measurement in the simplified model number k.

The simulated measurements of the virtual realities need to be perturbed by systematic and random measurement errors that reflect possible contributions to errors in field measurements (Barlebo et al., 2004). In particular, the uncertainty of placing the observations at the right place is accounted for by implementing a random perturbation $\varepsilon_{(p_i(d_j),l)}$ of the measurement locations $p_i(d_j)$ for each virtual reality $\mathbf{x}_0(l)$. Additionally, white noise $\hat{y}_0$ is added to each measurement at point i and virtual reality l drawn from a Gaussian distribution with zero mean and a standard deviation representing the standard measurement error. The latter corresponds to the resolution of the intended measurement device used in a field experiment.

In contrast to the virtual high-resolution realities, the simulated measurements based on the simplified model candidates for calibration are not perturbed: only the virtual realities have the job to mimic the outcomes of proposed field campaigns, while the simplified models will be calibrated against these mimicked outcomes.

For all virtual-reality realizations l and all conceptual models M_k, also the corresponding target quantities $t_0(l)$ and $t_k(l)$ are computed, which will be important for model selection in Step 5 (see Section 3.1.5).

3.1.4 Mimicking a Best-Estimate Calibration

Step 4 includes a best-estimate calibration for each virtual reality l and measurement design d_j, in which the simulated and perturbed data sets of the virtual realities are compared to measurement-error free predictions by the simplified conceptual models M_k. Rather than performing an iterative optimization, the existing ensembles of model candidates are used, each consisting of the same number n_r of realizations. For each virtual reality l, design d_j, and simplifying model M_k, the realization λ of the simplified model is determined with the smallest sum of squared differences $\phi(d_j, k, l, \lambda)$ between the perturbed simulated measurements of the virtual reality and non-perturbed simulated measurements of model M_k:

$$\phi(d_j, k, l, \lambda) = \sum_{i=1}^{n_{obs}(d_j)} \Big(y_0\big(p_i(d_j) + \varepsilon_{(p_i(d_j),l)}, l\big) + \tilde{y}_0 - y_k\big(p_i(d_j), \lambda\big) \Big)^2 \tag{3.1}$$

Note, that since I compare each virtual-reality realization l with all available realizations of a model candidate, I need to introduce the additional index λ. The best realization of model M_k is denoted $\lambda_{opt}(d_j, k, l)$:

$$\lambda_{opt}(d_j, k, l) = \underset{\lambda[1 \ldots n_r]}{\arg\min}\, \phi(d_j, k, l, \lambda) \tag{3.2}$$

This procedure is repeated with all virtual realities and all measurement designs, resulting in $n_m \times n_r$ best-fit models per measurement design. For the approach to actually mimic a best estimate calibration, the number of realizations n_r must be sufficiently large. Due to the measurement errors included in the virtual-reality runs and the inaccuracies of model simplifications, the index of the optimal realization $\lambda_{opt}(d_j, k, l)$ may differ from the index of the virtual reality l that has generated the data. Unlike in methods that aim at posterior distributions of parameters (e.g., Leube et al., 2012), l is not excluded from the set of candidates, as the peak of the likelihood in a real best-estimate calibration using the conceptual model M_k may be very close to the upscaled

parameter set $\mathbf{x}_k(l)$.
After having identified the best-fit pairs, the quality of the best fit is quantified with the root mean squared error $RMSE^y(d_j, k)$ for each design d_j and conceptual model M_k on average over all virtual-reality realizations l:

$$RMSE^y(d_j, k) = \sqrt{\frac{1}{n_{obs(d_j)}} \frac{1}{n_r} \sum_{l=1}^{n_r} \phi\big(d_j, k, l, \lambda = \lambda_{opt}(d_j, k, l)\big)} \tag{3.3}$$

In this thesis, the joint optimization framework is applied to a synthetic case study, in which the individual conceptual models M_k differ in the spatial resolution of the same parameter field x. Thus, one may expect that a model variant with a larger number of discretized parameters is better in meeting the perturbed measurements than a variant with fewer discretized parameters. That is, the corresponding error metric $\phi(d_j, k, l, \lambda_{opt}(d_j, k, l))$ should be smaller for k representing a model with more model parameters. However, the simulated measurements are prone to error, so that the higher flexibility of the model with more parameters may be wasted on reproducing these errors without increasing the prediction certainty – a phenomenon known as overfitting (Vanlier et al., 2014). Hence, to truly test which combination of conceptual model and measurement design is best, their ability to predict the target quantity t must be evaluated rather than their ability to reproduce the error affected measurements on which they have been calibrated.

3.1.5 Optimization with Respect to the Modeling Goal

After determining all best-fit models for each design and conceptual model, the performance of the selected models in predicting the target quantity t is analyzed, i.e. the modeling goal, over all virtual-reality realizations. Towards this end, the relative error $E^t_{rel}(d_j, k, l)$ is computed in the target quantity t for each virtual reality l, measurement design d_j, and conceptual model M_k:

$$E^t_{rel}(d_j, k, l) = \frac{t_0(l) - t_k\big(\lambda_{opt}(d_j, k, l)\big)}{t_0(l)} \tag{3.4}$$

With eq. (3.4) empirical distributions are obtained depending on the measurement design d_j and conceptual model M_k. These empirical distributions are analyzed by three

standard metrics: (1) the mean relative error $ME^t_{rel}(d_j, k)$, quantifying a bias in predicting the target quantity, (2) the standard deviation of the relative error $SDE^t_{rel}(d_j, k)$, quantifying the spread, and (3) the root mean squared relative error $RMSE^t_{rel}(d_j, k)$, quantifying the overall misfit:

$$ME^t_{rel}(d_j, k) = \frac{1}{n_r} \sum_{l=1}^{n_r} E^t_{rel}(d_j, k, l) \tag{3.5}$$

$$SDE^t_{rel}(d_j, k) = \sqrt{\frac{1}{n_r - 1} \sum_{l=1}^{n_r} \left(E^t_{rel}(d_j, k, l) - ME^t_{rel}(d_j, k) \right)^2} \tag{3.6}$$

$$RMSE^t_{rel}(d_j, k) = \sqrt{\frac{1}{n_r} \sum_{l=1}^{n_r} E^t_{rel}(d_j, k, l)^2} \tag{3.7}$$

The optimal combination of measurement design d_j and conceptual model M_k would have a mean relative error ME^t_{rel} of zero and a minimal spread SDE^t_{rel}, resulting in a minimal $RMSE^t_{rel}$. This yields the following criterion for jointly selecting the optimal measurement design d^t_{opt} and conceptual model M^t_{opt} considering the target quantity t:

$$\{d^t_{opt}, M^t_{opt}\} = \arg\min_{\{d_j, k\}} RMSE^t_{rel}(d_j, k) \tag{3.8}$$

Rather than jointly selecting the measurement design and conceptual model, the best design d_j could be determined for meeting a given target t, conditioned on a specific model selection M_k (or the best conceptual model conditioned on a specific design).

Chapter 4

Synthetic Case Study[1]

.

4.1 Application to Radial Flow in Stratified Aquifers

I investigate the hydraulic anisotropy in stratified aquifers using an approach similar to that of hydraulic tomography: I consider steady-shape pumping tests using partially screened extraction wells, with screens at various depths, and multi-level observation wells at various radial distances. For analyzing these data, a model needs to be specified considered as an appropriate representation of the complex subsurface. Calibrating a model with the attempt to fully resolve the **K**-field of an effectively heterogeneous subsurface would either demand an infeasible data amount or lead to enormous uncertainties. Simplifying the model requires important decisions on using a sequence of scalar, local conductivity values across some relevant number of strata, an effective macroscopic conductivity tensor over the entire aquifer formation, or something in between. Simultaneously, the challenge of designing the measurement strategy has to be addressed with which conclusive and sufficient data should be obtained while reducing the number of necessary observation wells to minimize costs. In facing this challenge, designing the data collection and choosing the affordable model complexity are inherently interconnected, and the optimal combination of measurement design and model selection depends on the ultimate application of the calibrated model.

[1] The essential content of Chapter 4 is published by Maier et al. (2020)

In the following, the optimization framework proposed in Chapter 3 is applied to a synthetic case study of radial flow in stratified aquifers. Therefore, hydraulic tests are considered as outlined in Section 2.2. The hydraulic tests were performed to guide the dimensioning of dewatering construction sites in fluvial aquifers that can be approximated as stratified systems. Both the hydraulic tests and the dewatering scenario are briefly described next, followed by an outline of Steps 2 to 5 of the optimization framework applied to the synthetic case study.

4.1.1 Hydraulic Tests

For the hydraulic tests steady-state groundwater flow was simulated to a partially penetrating well with water extraction from six well screens $i_{scr} \in \{1; 2; 3; 4; 5; 6\}$ placed at different depths. The setup of the hydraulic tests follows the governing equations and boundary conditions outlined in Section 2.2. As mentioned in Section 2.2, the steady-shape regime of the pumping tests was analyzed, i.e. the pumping regime in which drawdown differences between observation locations remain constant (Bohling et al., 2002, 2007).

Regarding the framework described in Section 3.1, the combined parameter fields K_r(z) and K_z(z) make up the parameter field x with different spatial resolutions among the conceptual models to be calibrated and in the virtual realities. As a consequence of the steady-shape pumping regime, the measured quantity y consists of drawdown-differences between piezometer screens rather than the absolute values. Quasi steady-state drawdowns were measured in several multi-level piezometers at different radial distances (Figure 4.1a).

4.1.2 Dewatering Scenario

As target application, the dewatering of a hypothetical cylindrical construction pit was considered and radial symmetric steady-state groundwater flow was simulated towards a construction pit within a large river. As shown in Figure 4.1b, the construction pit is separated from the surrounding river at the radial distance r_{sheet} by sheet piles extending to a depth z_{sheet} that is somewhat deeper than the bottom pit z_{pit}. Outside

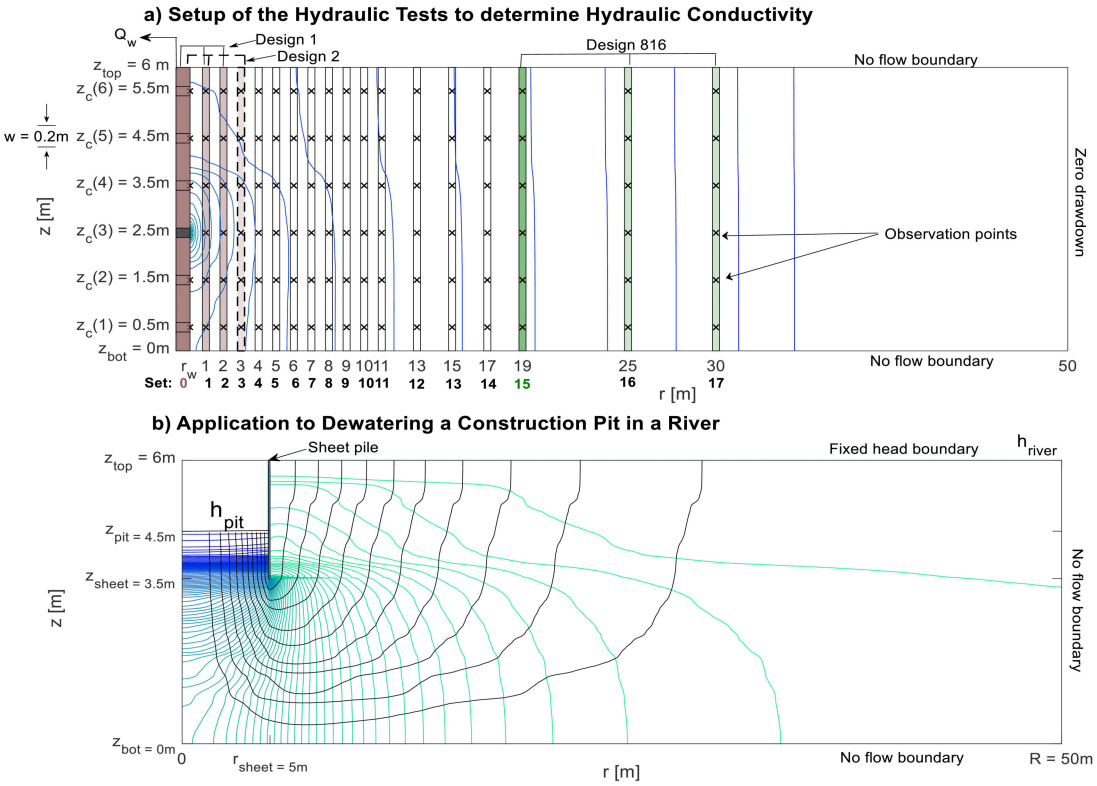

Figure 4.1: (a) Setup of the hydraulic test to estimate hydraulic conductivity: Radial symmetric steady state groundwater flow to a partially penetrating well. Water is successively extracted along six well-screen sections i_{scr} at $r = r_w = 0.1$ m and drawdown is observed at all depths of water extraction in 3 out of 18 radial distances to the pumping well. Analysis of 816 designs, each design combining three multi-level piezometers. (b) Application to the dewatering of a construction pit: Radial symmetric steady-state groundwater flow (colored lines: head contours; black lines: streamlines) towards a construction pit within a large river. The base of the construction pit is at z_{pit}, sheet piles pushed to depth $z = z_{sheet}$ are placed at $r = r_{sheet}$ to separate the construction pit from the surrounding river. (from Maier et al., 2020)

of the construction pit, the hydraulic head at the top equals the river stage h_{river} [L]. This implies the following fixed-head boundary conditions:

$$h(r < r_{sheet}, z = z_{pit}) = z_{pit} \tag{4.1}$$

$$h(r > r_{sheet}, z = z_{top}) = h_{river} \tag{4.2}$$

In addition, there are no-flow boundary conditions at the outer radial boundary of the domain, the bottom of the aquifer and along the sheet piles:

$$\left.\frac{\partial h}{\partial r}\right|_{r=R} = 0 \,\forall z \tag{4.3}$$

$$\left.\frac{\partial h}{\partial z}\right|_{z=z_{bot}} = 0 \,\forall r \tag{4.4}$$

$$\left.\frac{\partial h}{\partial r}\right|_{r=r_{sheet}, z_{sheet}<z<z_{top}} = 0 \tag{4.5}$$

The key target quantity to predict (t according to the nomenclature of Section 3.1.1) was defined as the dewatering flux Q_d needed to keep the water table in the pit at the bottom of the pit. This flux is computed by integrating the vertical specific discharge over the area of the construction pit:

$$Q_d = -K_z(z_{pit}) 2\pi \int_0^{r_{sheet}} r \left.\frac{dh}{dz}\right|_{z=z_{pit}} dr \tag{4.6}$$

As computing the dewatering flux required setting up and running an additional numerical model, I also tested whether the fully upscaled, anisotropic hydraulic conductivity tensor (see Section 2.3) is met by the estimated conductivity distribution resulting from the analysis of the hydraulic test. This resulted in two more target quantities, namely the effective horizontal and vertical conductivities K_r^{eff} and K_z^{eff}, respectively.

4.1.3 Proposed Designs of Observation Wells in the Hydraulic Test

To suggest potential measurement designs, spatial locations p_i need to be defined at which the drawdown may be measured. Trying to stay realistic regarding experimental feasibility, the number of multi-level piezometers was limited to three and the number of observation depths per multi-level piezometers to six, namely exactly at the depths of the screen sections of the extraction well. Figure 4.1a shows 18 potential distances at which the multi-level piezometer could be placed, including the multi-level extraction

well itself. Therefore, the total number of observation points p_i is $n_p = 108$ points. In

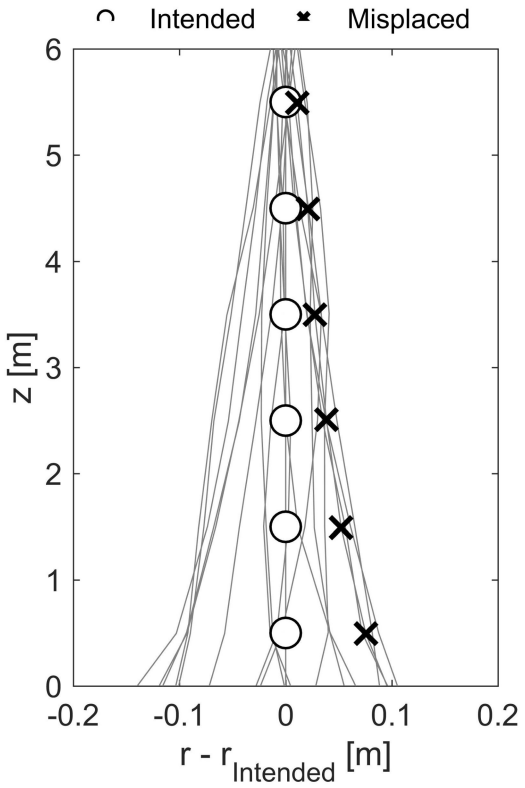

Figure 4.2: Misplacing piezometers by perturbing the r- and z-coordinate of an intended measurement point in the virtual realities. The variability of the horizontal placement increases with depth. (from Maier et al., 2020)

the virtual realities I considered that piezometers may be misplaced. This was simulated by considering erroneous r- and z-coordinates of an intended measurement point in the virtual realities (but not in the simplified models to be calibrated against the data). Towards this end, random values were computed drawn from uncorrelated uniform distributions on the interval [-0.025m; 0.025m] and [-0.02m; 0.02m] for perturbing the r- and z-coordinates of a proposed observation point, respectively. In the field installation of piezometers the horizontal position may increasingly drift with depth, causing increasing uncertainty of the actual horizontal position with increasing depth. To account for this uncertainty, the r-coordinates of measurement points were additionally

perturbed by the following double-cumulative noise term:

$$\varepsilon_{r,cum}(z_i) = 5 \times 10^{-5}[m] \sum_{j=1}^{i} \sum_{k=1}^{j} \omega_k \tag{4.7}$$

in which the summation is performed over the layers of the numerical grid starting from the top, z_i is the vertical coordinate of i-th grid layer from the top, and ω_j is drawn from a standard normal distribution. Information on the spatial discretization is given in Section 4.1.4.

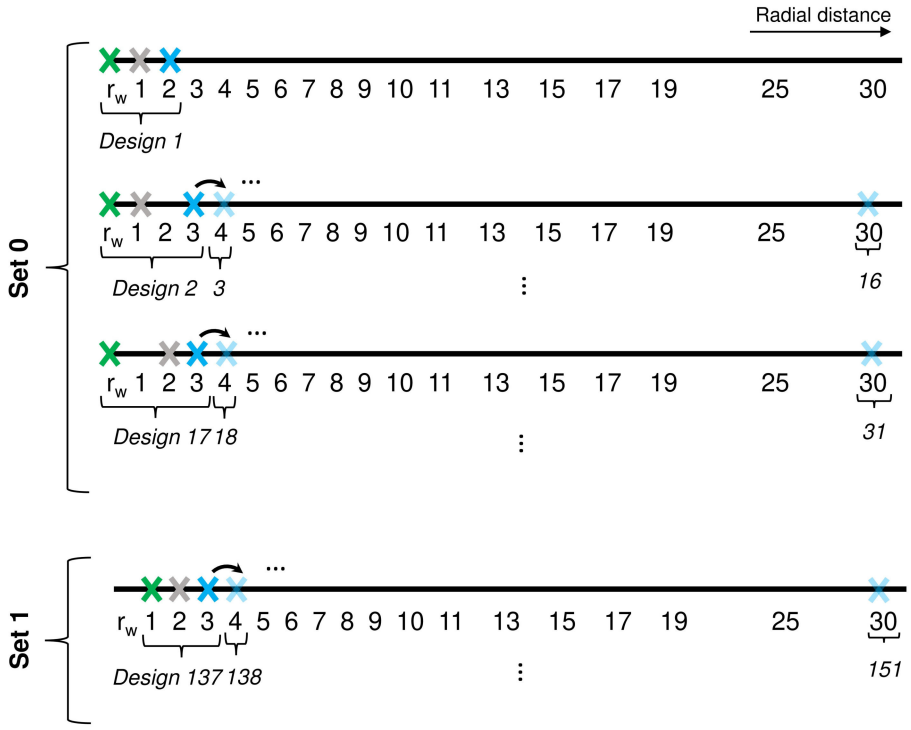

Figure 4.3: Setup and sorting of designs by means of set 0 and set 1. In each set, the piezometer closest to the pumping well (green cross) is considered as fixed. The two remaining piezometers are shifted further outwards, starting with the outermost one (blue cross). After the outermost piezometer has reached the last radial distance available, the second piezometer (grey cross) is moved to the next position. This procedure is repeated until the second and third piezometer have reached the two last available positions. This is when the respective set is completed and the next set starts by moving the piezometer closest to the pumping well to the next position and so forth. (from Maier et al., 2020)

An example of resulting deviations in intended and misplaced measurement locations is shown in Figure 4.2. Since the pumping well was not considered to be affected from an imprecise placement of measurement location such errors were excluded in the therein considered observation points.

Considering drawdown differences, each design includes $3 \times 6 - 1 = 17$ observations. However, for designs involving the extraction well the pumped screen section itself was excluded as an observation point since in field applications the extraction screen is typically influenced by well-skin effects and pressure fluctuations caused by the pump. Thus, such designs only include 16 drawdown differences. There are in total $n_d = 816$ possibilities to combine three out of 18 multi-level piezometers, making up the individual measurement designs d_j. The designs are grouped into sets numbered 0 to 15 according to the distance of the respective observation well closest to the pumping well (Figure 4.1a). Within one set, the designs are sorted by shifting the observation wells further outwards, starting with the outermost well until it reaches the last possible radial distance (Figure 4.3).

In each design d_j, the lowermost observation point of the observation well closest to the pumping well constitutes the reference point for generating drawdown differences Δs. For the particular case that water is extracted at the lowermost depth when analyzing designs of set 0, the reference point is changed to the second lowermost observation point for generating Δs. Each simulated drawdown of the virtual realities was independently perturbed by a measurement error ε_s drawn from a normal distribution with zero mean and standard deviation $\sigma_s = 3$ mm corresponding to the resolution of the intended measurement device in a field test.

4.1.4 Generating Realizations of Hydraulic Conductivity as Virtual Realities and for Model Calibration

The ensemble of virtual realities with the parameter sets $\mathbf{x}_0$ contains $n_r = 20,000$ realizations of highly resolved vertical profiles of hydraulic conductivity. Towards this end, multi Gaussian one-dimensional fields of the horizontal log-hydraulic conductivity were generated with an exponential covariance function of uncertain mean, variance, and correlation length. For each realization the mean, variance, and correlation length were drawn from a uniform distribution on the interval of the respective minimum and

maximum (Table 4.1b) before generating the autocorrelated vertical profile of $\ln K_r(z)$. Figure 5a shows the example of a highly resolved conductivity profile with the blue line marking the local horizontal conductivity and the red line the slightly smaller vertical conductivity. To obtain local values of the vertical hydraulic conductivity, a random local anisotropy factor between 1 and 2 was employed using the cumulative distribution function of a standard normal distribution shifted by 1. Each highly resolved

Table 4.1: Parameters of the tests case.

a. Pumping rate in hydraulic tests		
	Hydraulic tests	**Dewatering scenario**
Pumping rate Q_w	0.004 m^3/s	-
b. Ranges of statistical hyper-parameters of lnK		
	Minimum	**Maximum**
Mean	-11.5	-4.6
Variance	0.5	3
Correlation length	0.1m	1.5m
c. Spatial discretization		
	Hydraulic tests	**Dewatering scenario**
Vertical discretization	0.025m	0.025m
Radial discretization	Logarithmically increasing from well radius r_w = 0.1m to radius R = 50m	0.1m for r<r_{sheet} = 5m; Logarithmically increasing from r_{sheet} to radius R = 50m

hydraulic conductivity field, representing a virtual reality, was partially upscaled to obtain the candidate parameter fields of the $n_m \times n_r$ realizations of simplified models. The first conceptual model assumes a single homogeneous anisotropic layer, denoted 1-layer model. The second model assumes two layers of equal thickness, the third model three layers and the fourth model six layers, denoted 2-, 3-, and 6-layer model, respectively. The partial upscaling of the highly resolved K_r-field and K_z-field were performed according to eq. (2.10) and eq. (2.11), respectively.

Figure 4.4b illustrates the partially upscaled conductivity profiles according to the 1-, 2-, 3- and 6-layer model derived from the highly resolved profile shown in Figure 4.4a. K_r^{eff} and K_z^{eff} of the 1-layer model represent the fully upscaled radial and vertical conductivities and act both as a candidate in the calibration exercise and as alternative targets t of the overall optimization.

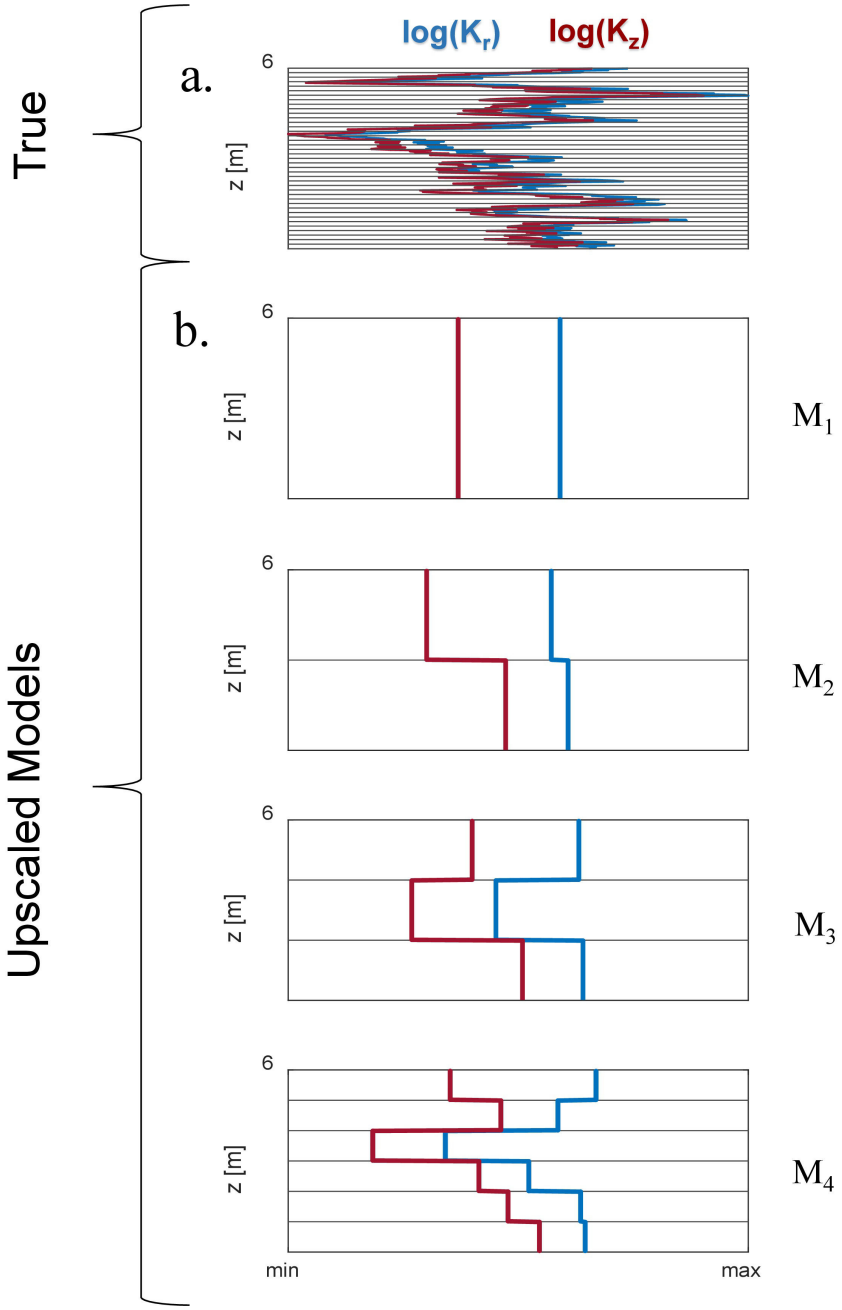

Figure 4.4: Example of vertical profiles of horizontal and vertical hydraulic conductivity, K_r and K_z, respectively. (a) Highly resolved profile used as virtual reality; (b) partially upscaled profiles of K_r and K_z according to the models of 1, 2, 3, and 6 horizontal layers used as candidate fields in the calibration-mimicry. (from Maier et al., 2020)

4.1.5 Numerical Implementation and Computational Effort

The entire computational code was written in Matlab. The generation of highly-resolved auto-correlated hydraulic-conductivity fields was done by the spectral approach of Dietrich and Newsam (1993), whereas the flow fields were computed by conforming Finite Elements using bilinear elements on a rectangular grid.

With the ensemble size of $n_r = 20{,}000$ realizations, $6 \times 20{,}000 = 120{,}000$ drawdown fields of the hydraulic tests had to be simulated using the virtual truths, one for each of the six extraction depths and for each realization. From this, all hypothetical measurement values for all 816 designs and 20,000 realizations could be constructed. Also, 20,000 simulations of the dewatering example had to be performed using the virtual realities in order to obtain the correct values of the target quantity for Step 5. Information on the discretization of both numerical models is given in Table 4.1c.
The computational effort for the simulations using the simplified conceptual models was four times larger because four conceptual models were tested.

All other steps followed the procedure outlined in Section 3.1:

- The locations of all potential observation points were perturbed in all realizations of the virtual truth. The simulated drawdown values at these locations were then picked and perturbed with measurement noise. The right drawdown differences were assigned to the right measurement designs to obtain the virtual measurements $y_0(p_i(d_j) + \varepsilon_{p_i(d_j),l}, l + \varepsilon_{y_0(p_i(d_j),l)}$ for all realizations l of the virtual truth and all designs d_j (part of Step 3 in Figure 3.1).

- The target quantities $K_r^{eff}(l)$, $K_z^{eff}(l)$, and $Q_d(l)$ were computed for all realizations (l) of the virtual truth (part of Step 3 in Figure 3.1).

- The simulated drawdown values were picked at all unperturbed potential observation points in all realizations of all four candidate models and the right drawdown differences were assigned to the right measurement designs to obtain the simulated measurements $y_k(p_i(d_j), \lambda)$ of the different candidate models (part of Step 3 in Figure 3.1).

- The best-estimate calibration was mimicked for each virtual reality l, measurement design d_j and conceptual model M_k according to equation 3.2 (Step 4 in Figure 3.1).

- The combination of measurement design d_j for a given conceptual model M_k was determined (or the best combination of the latter two) by finding the minimum of the root mean squared relative errors $RMSE^t_{rel}(d_j, k)$ according to equation 3.7 with t being K_r^{eff}, K_z^{eff}, or Q_d (Step 5 in Figure 3.1).

The overall computational effort on a Fujitsu P957 / power High End PC was 336.1 hours wall-clock time.

4.2 Results and Discussion

4.2.1 Reproduction of Measurements by the Calibrated Models

Figure 4.5 illustrates the root mean square error $RMSE^y(d_j, k)$ (see eq. (3.3)) of the best-fit realizations λ_{opt} for each measurement design d_j and conceptual-model candidate k as colored lines (red: 1-layer model, green: 2-layer model, blue; 3-layer model, gold: 6-layer model). The $RMSE^y(d_j, k)$-values quantify how well the measurements simulated by the calibrated simplified models agree with the simulated and perturbed measurements of the virtual realities after mimicking the calibration (Step 4). For comparison Figure 4.5 also contains a black line representing the root mean squared value of the measurements in the virtual realities quantifying the total strength of the signal. Because Figure 4.5 shows the $RMSE^y(d_j, k)$-values on a semi-logarithmic scale, the distance between the colored and the black line can be interpreted as relative error of the calibration. For additional comparison, the mean measurement error considered in the field tests was computed (purple line in Figure 4.5).

First, a discussion of the general trends is provided: As described in Section 4.1.3, the designs are sorted into sets comprising all designs with the same observation well closest to the pumping well. Set 0 includes the non-pumped well screens of the pumping well as observation points (see Figure 4.1a). As seen in Figure 4.5, the $RMSE^y(d_j, k)$-values in general decrease with increasing distance between the pumping well and the first multi-level piezometer and approach the measurement error (see purple line of Figure 4.5). However, also the measured drawdown-differences decrease (see black line of Figure 4.5) so that the relative misfits for a given conceptual model do not differ so strongly between the different design sets.

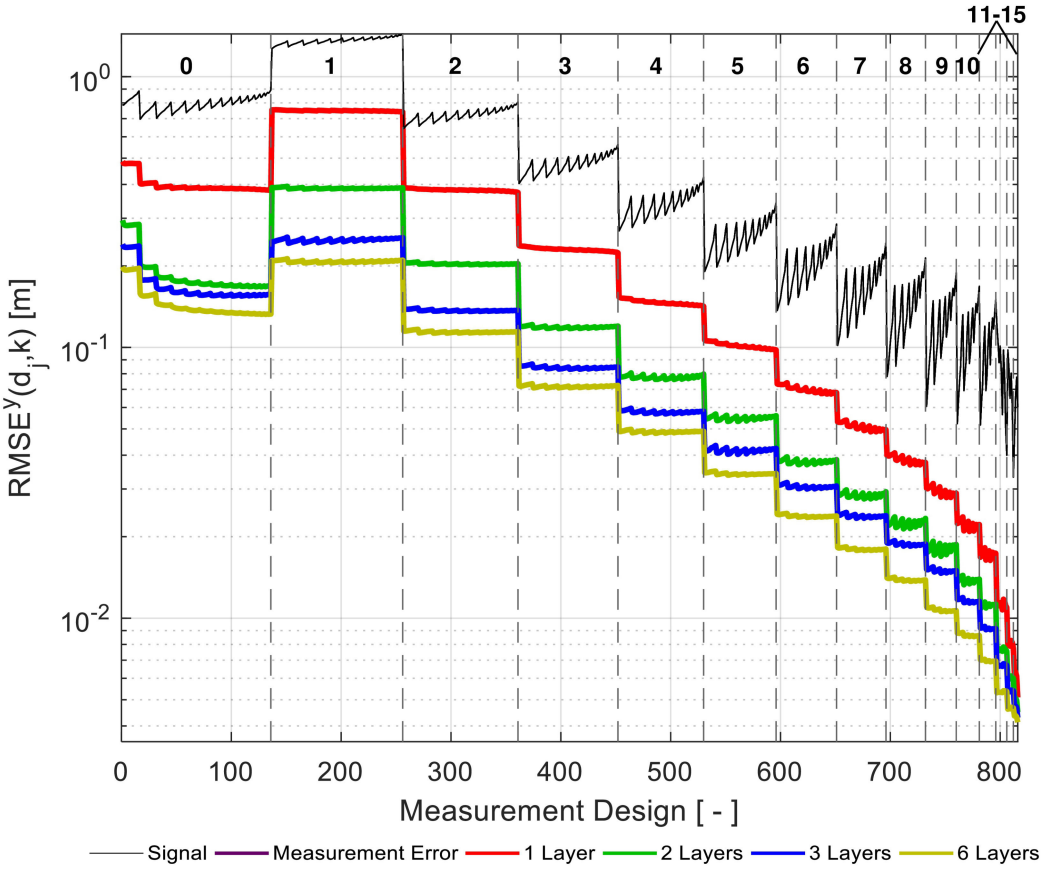

Figure 4.5: Root mean square error $RMSE^y(d_j, k)$ of the best-fit of the 1-, 2-, 3-, and 6-layer models in meeting the drawdown differences of the virtual reality as function of the measurement design. Black line: root mean squared values of the measurements. Purple line: Measurement error expected from the measurement devices employed in field tests. (from Maier et al., 2020)

As can be seen by the illustrative drawdown contour-lines in Figure 4.1a, the vertical variation in drawdown decreases with increasing distance to the pumping well so that a distinction between models with many or few layers becomes hardly possible when only distant multi-level piezometers are considered. For the same reason, the imprecise location of the observation points in the virtual realities has a higher effect on the absolute and relative measurement error for close-by than for far-away piezometers. However, such misplacements were not considered for observations at the non-pumped well screens of the pumping well. The latter may explain why, contrary to the general trend, the design set 1 results in larger calibration errors than the design set 0.

In a next step, the models are compared: Obviously, for all designs, the 6-layer model shows the smallest misfit of the measurements. This confirms the trivial expectation that a model with a higher spatial resolution of the hydraulic conductivity can reproduce measurements better than a model with a coarser resolution. In the design set including the extraction well as a piezometer, the 2-, and 3-layer models perform similarly well as the 6-layer model, whereas the $RMSE^y(d_j, k)$-curves of the model candidates separate when considering designs of larger set numbers. This also accounts for the 1-layer model which performs worst for all designs, followed by the 2- and 3-layer models. The increased spreading between model candidates with increased set number (pronounced by the logarithmic scaling in Figure 4.5) results from the decrease in signal (see black line in Figure 4.5) suggesting that increasing the spatial resolution of a model for reproducing measurements especially matters when measurements are afflicted with a small signal strength.

4.2.2 Predictive Performance of Calibrated Models

While the 6-layer model is best in reproducing the observed drawdown-differences, this does not guarantee that it is also best in predicting the desired target quantities, as there is the danger of overfitting. To assess the predictive performance of the models, the relative error was computed in predicting the target quantities of the dewatering flux Q_d, the effective radial hydraulic conductivity K_r^{eff} upscaled over the full aquifer thickness, and the equally effective vertical hydraulic conductivity K_z^{eff}.

First, the empirical probability is reported that a particular calibrated model M_k performs best in meeting an individual target quantity t for a given design d_j over all realizations $l = 1 \ldots n_r$ of the virtual reality. Per individual realization l, the selection of best-performing model is denoted $k_{opt}^t(d_j, l)$:

$$k_{opt}^t(d_j, l) = \arg\min_k E_{rel}^t(d_j, k, l)^2 \tag{4.8}$$

Then, the relative frequencies of model selection are counted across the virtual-reality realizations. Figure 7 shows the resulting selection probability for the 1-, 2-, 3-, or 6-layer models, respectively. Each subplot is related to one of the three target quantities (dewatering flux Q_d and fully upscaled horizontal and vertical hydraulic conductivities K_r^{eff} and K_z^{eff}).

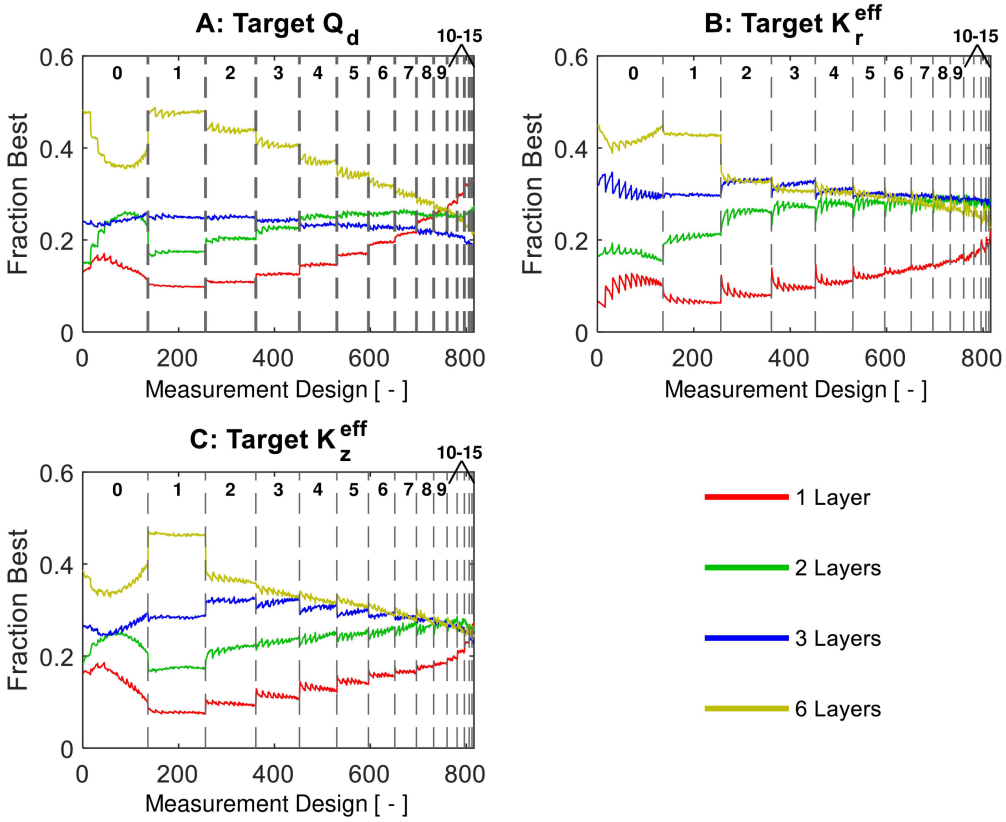

Figure 4.6: Fraction of all realizations n_r in which a particular model candidate was determined as best among the available model candidates in minimizing $E^t_{rel}(d_j, k, l)^2$ of the target quantities: (a) Q_d, (b) K_r^{eff} and (c) K_z^{eff}. (from Maier et al., 2020)

Looking at general trends across all target quantities shows that all selection probabilities range between 5.4% and 49%. This indicates that none of the model candidates explicitly predominates model selection. In general, the selection probability of the 1-, 2-, 3-, and 6-layer models reach similar values with increasing design number, that is, with increasing distance between the pumping well and the observation wells. As already discussed in Section 4.2.1, the vertical variation in drawdown decreases with increasing distance to the pumping well. This leads to a better fit in calibration. In the current context, this also means that a distinction between models becomes hardly possible when only distant multi level piezometers are considered.

Looking at individual model errors, for all target quantities the homogeneous 1-layer model is selected the least. This even holds for the fully upscaled directional hydraulic

conductivities K_r^{eff} and K_z^{eff}, which are uniform effective parameters that should relate perfectly to the 1-layer conceptual model (per definition, see Section 2.3). This may be explained with the discrepancy by the spatially non-uniform sensitivities of the drawdown measurements with respect to hydraulic conductivity even under uniform hydraulic conditions, which leads to a bias in the estimated conductivities of the 1-layer model. Since the 1-layer model considers the largest possible scale, its selection probability for the fully upscaled directional hydraulic conductivities K_r^{eff} and K_z^{eff} improves for piezometers at far radial distances.

In a next step, the individual target quantities are discussed. Figure 4.6a shows that the 6-layer model has the highest fraction of best predicting the dewatering flux Q_d among the model candidates, regardless of the measurement designs in set 0-10. This may be explained by the flow field in the vicinity of the construction pit shown in Figure 4.1b: The dewatering flux crucially depends on (a) the hydraulic conductivity tensor around the bottom end of the sheet pile and (b) the vertical hydraulic conductivity integrated between the bottom of the sheet pile z_{sheet} and the bottom of the construction pit z_{pit}. The 6-layer model is suited best to capture these properties.

In contrast, for predicting K_r^{eff} and K_z^{eff}, the 6-layer model is not chosen as frequently as the best model (Figure 4.6b and c). Here, first the 3-layer model and then the 2-layer model approximate the best performing model (predicting K_z^{eff}) or perform just as well as the 6-layer model (predicting K_r^{eff}) when increasing the distance of the multi-level piezometers to the pumping well. This can be explained with the effect of overfitting the erroneous measurements in the virtual reality suggesting that the model complexity of a 6-layer model is required for predicting Q_d, yet does not necessarily yield better results for predicting K_r^{eff} and K_z^{eff}.

4.2.3 Predictive Errors

Finally, to provide quantitative statements, Figure 4.7 shows the root mean squared relative errors $RMSE_{rel}^t$ in predicting the three target quantities in all 20,000 realizations considering the 816 different designs with the calibrated 1-, 2-, 3-, and 6-layer models. The individual subplots refer to the chosen conceptual models, whereas the line colors distinguish the target quantities. For all designs and models, the $RMSE_{rel}^t$-values strongly depend on the targeted parameter K_r^{eff}, K_z^{eff}, or Q_d, respectively. The

errors vary between 0.13 and 3.66, indicating that the target quantities are under- or overestimated by 13% to 366%, depending on the selected measurement design, model candidate, and target quantity. For most measurement designs and model candidates, predicting the fully upscaled horizontal hydraulic conductivity K_r^{eff} is easier than predicting the fully-upscaled vertical conductivity K_z^{eff} and the dewatering flux Q_d. This can be explained by the flow being predominantly horizontal, so that the designs are more informative about horizontal properties of the system. Over all designs, the av-

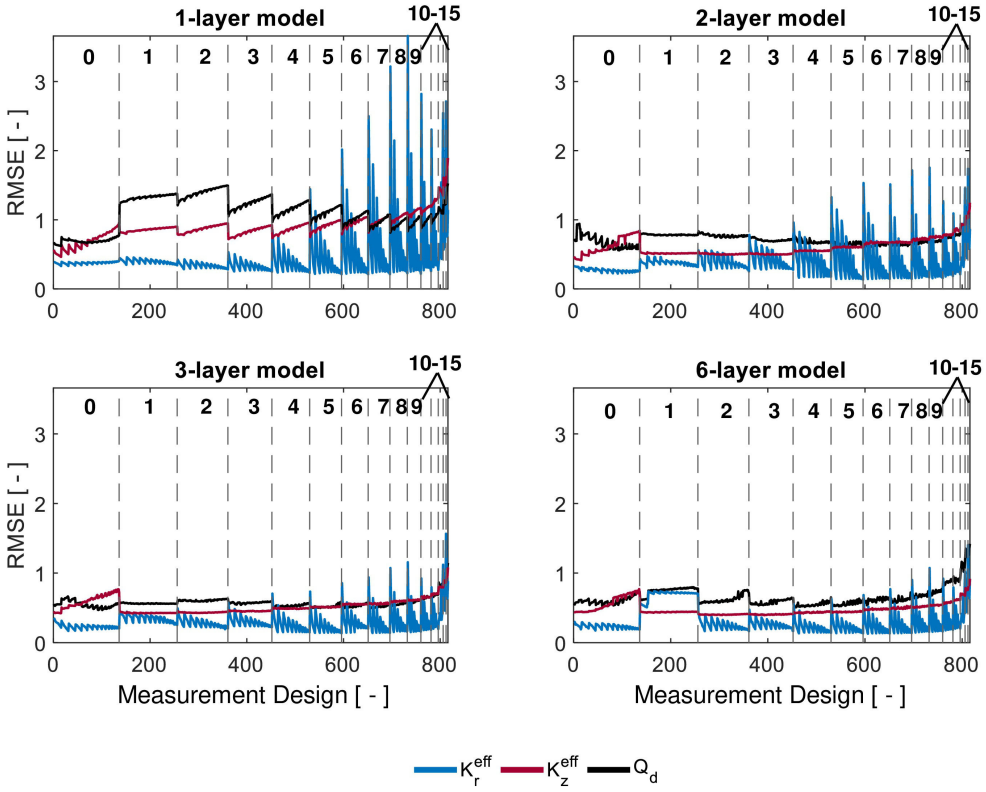

Figure 4.7: Root mean squared relative errors $RMSE_{rel}^t$ in predicting K_r^{eff} (blue), K_z^{eff} (red), and Q_d (black) as function of the chosen measurement designs for individual model choices. (a) 1-layer model, (b) 2-layer model, (c) 3-layer model, (d) 6-layer model. (from Maier et al., 2020)

erage magnitude of $RMSE_{rel}^t$ depends on the selected model, with the 1-layer model exhibiting the largest offsets between the correct target quantity values and the prediction by the calibrated model, whereas the most complex 6-layer model scores mostly

the best, even though the prediction errors of the 3-layer model highly resemble those of the 6-layer model.

The $RMSE^t_{rel}$-values show characteristic trends among the analyzed measurement designs, which are similar among the model candidates but appear to be differently pronounced due to the given dependence on systematic model errors. In the following, these trends are discussed for the different target quantities.

In general, the highest error in predicting K_r^{eff} (blue lines in Figure 4.7a-d) is obtained by applying the first design within each set. These are the designs in which the three multi-level piezometers are the closest to each other. With increasing design number within a set, the errors in predicting K_r^{eff} fluctuate showing a decreasing trend towards the last design of the associated set. In order to predict the fully upscaled horizontal hydraulic conductivity K_r^{eff} it is best to include measurements with the largest horizontal offset possible, and in a given set this is achieved by taking the two outermost positions for piezometers two and three. However, it is neither preferable to choose a first piezometer location that is too close to the pumping well, where the effects of misplacing the piezometers are the biggest, nor too far away where the measured signal gets too small. This implies that the least uncertainty in predicting K_r^{eff} is achieved with the first multi-level piezometer being placed at intermediate distances to the extraction well, and the two others as far away as possible.

In contrast to the $RMSE^t_{rel}$-values for K_r^{eff}, those for K_z^{eff} and Q_d show an increasing trend within each set (red and black lines in Figure 4.7a-d), expressing that the prediction uncertainty is reduced when the second multi-level piezometer is placed close to the first rather than the third one. This can be explained by the necessity of strong vertical drawdown differences in the observations to obtain good estimates of the vertical hydraulic conductivity, and strong vertical gradients are found closer to the extraction well. Similar to the error in predicting K_r^{eff}, the measurement-design sets with the first observation well at a close, but not too close distance to the pumping well seem to guarantee the smallest prediction error of K_z^{eff} and Q_d.

4.2.4 Multi-Objective Optimization

Equation (3.8) describes how to select the single best combination of measurement design d_j and conceptual model M_k to be calibrated to the data in order to meet an

individual target quantity $t \in \{Q_d, K_r^{eff}, K_z^{eff}\}$. Depending on the target, the results reveal different optimal designs that are visualized in Figure 4.8a. In order to satisfy all three criteria, a multi-objective optimization is needed (resulting designs see Figure 4.8b).

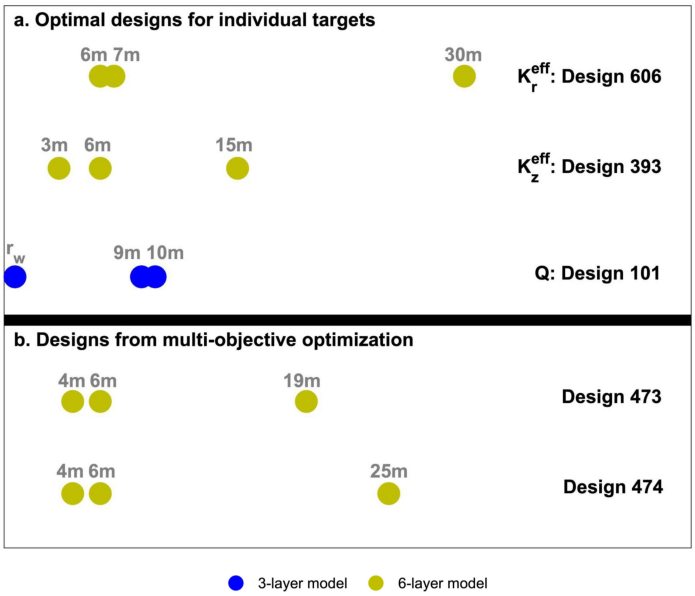

Figure 4.8: Optimized modeling and measurement strategies for reducing the relative prediction error of target quantities. The color of the points refers to the resulting model candidate, i.e., gold for the 6-layer model and blue for the 3-layer model. a. Results for minimizing errors of each target quantity K_r^{eff}, K_z^{eff}, and Q_d, respectively. Choosing one of these designs and models implies making a compromise with respect to minimizing the prediction uncertainty of the respectively two other target quantities. b. Designs resulting from multi-objective optimization. (from Maier et al., 2020)

For this purpose, potential trade offs were examined in predicting the different target quantities with a certain model and measurement strategy following the Pareto principle. Towards this end, the $RMSE_{rel}^t$-values were considered for all target quantities from all available combinations $n_m \times n_d = 3'264$ of conceptual models and measurement designs and are plot against each other in Figure 4.9. In particular, Figure 4.9a-c show pairwise Pareto plots of the relative prediction error in two of the three target quantities (Figure 4.9a: K_r^{eff} and K_z^{eff}; Figure 4.9b: K_r^{eff} and Q_d; Figure 4.9c: K_z^{eff} and Q_d) for all combinations of calibrated model candidates and measurement designs, whereas Figure 4.9d shows a three-dimensional Pareto plot of all three $RMSE_{rel}^t$-values. The

red, green, blue, and gold colors of points and spheres in Figure 4.9 relate to the 1-, 2-, 3-, and 6-layer models, respectively.

Figure 4.9 illustrates that there are no big trade-off effects when optimizing the model-and-design combinations for the fully upscaled horizontal conductivity K_r^{eff} and either the corresponding fully-upscaled vertical conductivity K_z^{eff} (Figure 4.9a) or the dewatering flux Q_d (Figure 4.9b). Meeting K_z^{eff} and Q_d clearly leads to a preference for the 6-layer model and 3-layer model, respectively, whereas meeting K_r^{eff} alone does not show such a preference. However, meeting the criterion of minimizing $RMSE_{rel}^t$ for $t = K_r^{eff}$ does not truly exclude meeting the other two criteria. The Pareto front (not shown) was computed, but it contained only a very small number of points, confirming the lacking trade-off between meeting K_r^{eff} and K_z^{eff} and K_r^{eff} and Q_d, respectively. As

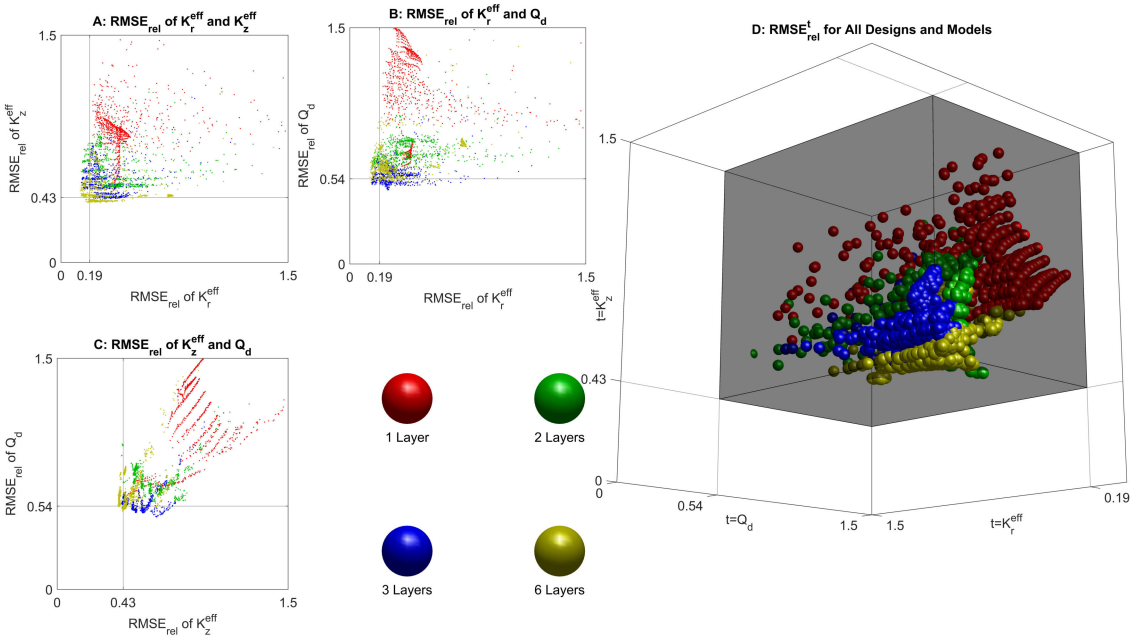

Figure 4.9: Root mean squared relative errors $RMSE_{rel}^t$ in predicting the three target quantities for all measurement designs and all model choices (indicated by colors). (a) pairwise plot of $RMSE_{rel}^t$ for $t = K_r^{eff}$ and $t = K_z^{eff}$, (b) pairwise for $t = K_r^{eff}$ and $t = Q_d$, (c) pairwise for $t = K_z^{eff}$ and $t = Q_d$, (d) $RMSE_{rel}^t$-values for all three targets. Gray lines in (a)-(c) and planes in (d): separation of the 10% best from the rest. (from Maier et al., 2020)

indicated by Figure 4.9a and Figure 4.9b, Figure 4.9c shows that the 6-layer model is the preferred model for minimizing $RMSE_{rel}^t$ for $t = K_z^{eff}$ whereas for minimizing $RMSE_{rel}^t$

for $t = Q_d$ the 3-layer model pre-dominates model selection. However, similar to their lacking trade-offs with K_r^{eff}, selecting a model-and-design combination for predicting the target K_z^{eff} does not necessarily deteriorate the prediction of Q_d.

Instead of computing the Pareto front, I thus decided to set threshold values corresponding to the 10% best model-and-design combinations of the 3'264 choices ($RMSE_{rel}^t$ = 0.19 for $t = K_r^{eff}$, $RMSE_{rel}^t = 0.43$ for $t = K_z^{eff}$, and $RMSE_{rel}^t = 0.54$ for $t = Q_d$). Figure 4.9a-c show these threshold values as gray lines, and Figure 4.9d as translucent gray planes parallel to the axes. The spheres with clear colors in the foreground of Figure 4.9d belong at least in one of the three criteria to the 10% best model-and-design combinations, and those of them closest to the observer belong to the best regarding multiple criteria.

There are exactly two model-and-design combinations that belong to the 10% best with respect to K_r^{eff}, K_z^{eff}, and Q_d. Both suggest to analyze the data by the 6-layer model and to install the closest multi-level piezometer at 4 m distance to the pumping well, and the second closest at 6 m, whereas the location of the furthest multi-level piezometer is chosen either at 19 m or at 25 m distance (see also Figure 4.8b). I consider these two designs as the overall best ones.

Chapter 5

Field Application

The synthetic case study presented in Chapter 4 has shown that calibrating groundwater-flow models using data from hydraulic tests with a tomographic layout is a promising method for estimating the large-scale hydraulic anisotropy. To test a field application of the method, I investigated a fluvial gravel aquifer in South-West Germany. A new field site was instrumented with a measurement network that was especially designed for performing the hydraulic tests at the site. In the center of the well network, a large-diameter well, having three separate well-screen sections, was installed while 14 nested observation wells were placed in four different directions from the pumping well. To facilitate depth-oriented observations, all nested observation wells were installed with screen openings in different depth, yielding a total of 58 observation points.

At the field site, a total of 22 pumping tests were performed in which water was successively extracted from the three well-screen sections of the pumping well. During each pumping test, the hydraulic response was measured in all available observation points. According to the synthetic case study (see Chapter 4), the field measurements were analyzed considering the steady-shape pumping regime of the tests. All raw and processed data are available on the FDAT repository of the University of Tübingen (http://hdl.handle.net/10900.1/274eabc9-0edd-40b1-9e34-541c09c4479c, Maier et al., 2021, available from January 2022).

The hydraulic tests were simulated and the hydraulic anisotropy was estimated by calibrating two groundwater flow models with the field measurements. Similar to the synthetic case study, a homogeneous anisotropic groundwater-flow model and a

groundwater-flow model, consisting of multiple, locally anisotropic, horizontal layers, were considered.

In the following, I describe the newly instrumented field site, delineate the pumping tests performed at the site and present the different steps considered in data processing. I then present the numerical groundwater-flow models, the calibration strategy, and the model-calibration results.

5.1 Field Site and Well Instrumentation

The field site is located in the Upper Rhine Valley, north of the municipality Kappel-Grafenhausen, Baden-Württemberg, South-West Germany (Figure 5.1a). The field site is surrounded by a diversion channel for flood protection in the west and by River Elz in the east. The general groundwater flow direction is from the southeast to the northwest with a hydraulic gradient of 0.14%. The fluvial unconsolidated aquifer consists of Quaternary sediments of the Neuenburg-Formation (qNE), and is characterized by a stratification of fine to sandy gravels and sands (LGRB, 2004; Wirsing and Luz, 2007). The saturated aquifer of $\sim$41 m thickness is overlain by a $\sim$2 m thick layer of alluvial fines and bounded by considerably less conductive Quaternary sediments from the Breisgau-Formation (qBR) at the aquifer base (LGRB, 2004; Wirsing and Luz, 2007). Results from a grain size analysis of an exploration drilling at the field site indicate hydraulic-conductivity values of the aquifer in the range of 6.7×10^{-5} m/s to 2.6×10^{-1} m/s. In further preliminary hydrogeological investigations at the site, eleven direct-push injection logs were performed at different locations. The results reveal no significant lateral differences in hydraulic conductivity but indicate lower hydraulic conductivities at depths from 5-10 m with respect to the aquifer top.

Figure 5.1b shows a plan view of the field installation at the site. A large pumping well (denoted R01) was installed via grab drilling in the center of the well network (Figure 5.2). The pumping well has a diameter of 800 mm and reaches to a depth of 21 m within the aquifer. It was designed with three separate screen sections, denoted I, II, and III, of 2 m length each and with a spacing of 4.5 m in between, centered at elevations of 37.32 m (upper screen), 30.82 m (middle screen) and 24.32 m (lower screen) above the aquifer base. Figure 5.3 shows a schematic 3-D illustration of the

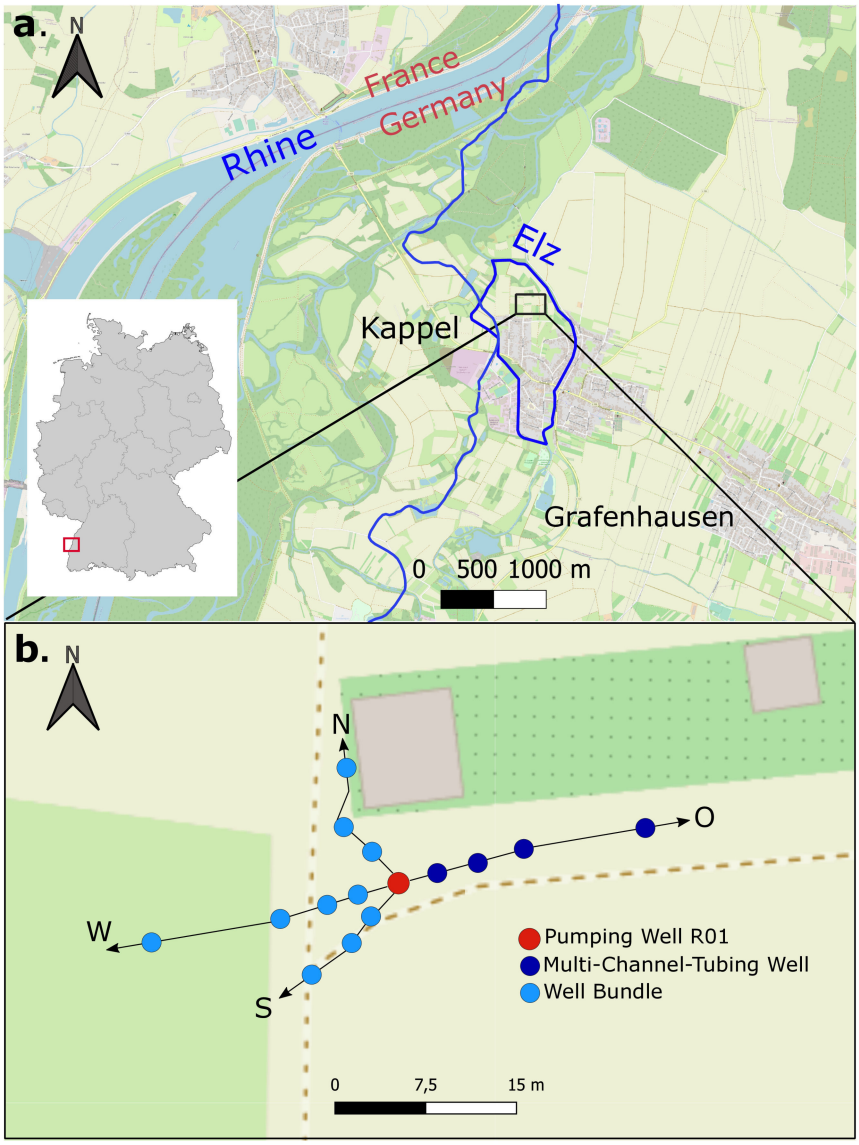

Figure 5.1: a. Overview map of the field site. b. Plan view of the installed measurement network at the site.

well configuration. Note that all z-coordinates are given in reference to the aquifer bottom. The well was installed with a prepacked filter along the filter-screen sections and completed with coarse gravel to a 0.4 m-thick filter pack, extending by 0.5 m above and below the individual screen sections. Each filter-pack section is connected above and below to a clay fill through a 0.5 m thick secondary filter layer. Three groups of

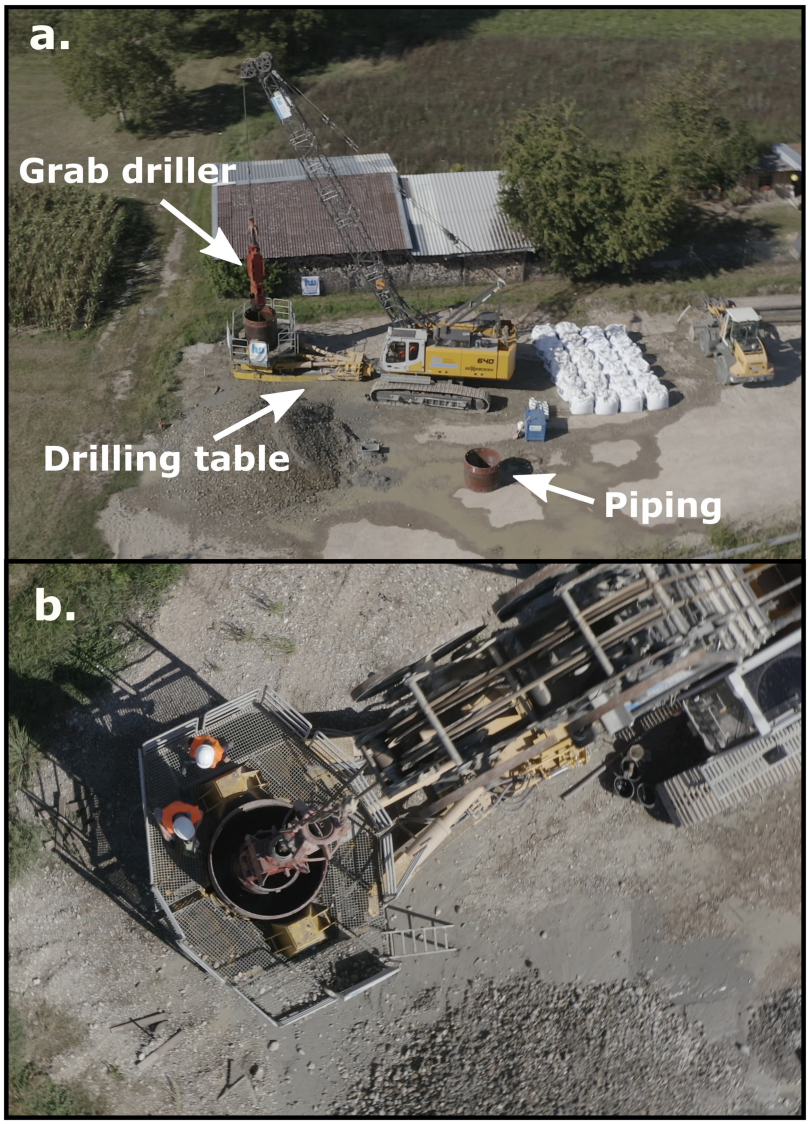

Figure 5.2: a. Side view and b. plan view of the grab drilling of pumping well R01.

nested observation wells were installed each in the north, east, south, and west direction of the pumping well R01, at radial distances of $r_1 \approx 3.5$ m, $r_2 \approx 6.5$ m and $r_3 \approx 10.5$ m, respectively (Figure 5.3). Two additional groups of nested observation wells were installed at a radial distance of $r_4 \approx 21$ m, one to the east and the other to the west of well R01. To facilitate a depth-oriented observation, each of the 14 groups of nested observation wells provides screen openings at different depths. The design of the nested observation wells is described in more detail in the following.

Each group of nested observation wells positioned north, south and west of R01 comprises three 1-inch piezometers, denoted three-well-bundle, that were installed with direct-push drilling and have a 0.3 m long screen at its bottom. Two of the three piezometers are placed at depths between the elevation of the pumping-well screens I and II, and the third is at a depth between the elevation of the pumping-well screens II and III (Figure 5.3). The individual piezometers were completed by allowing natural collaps to occur in the annular well space.

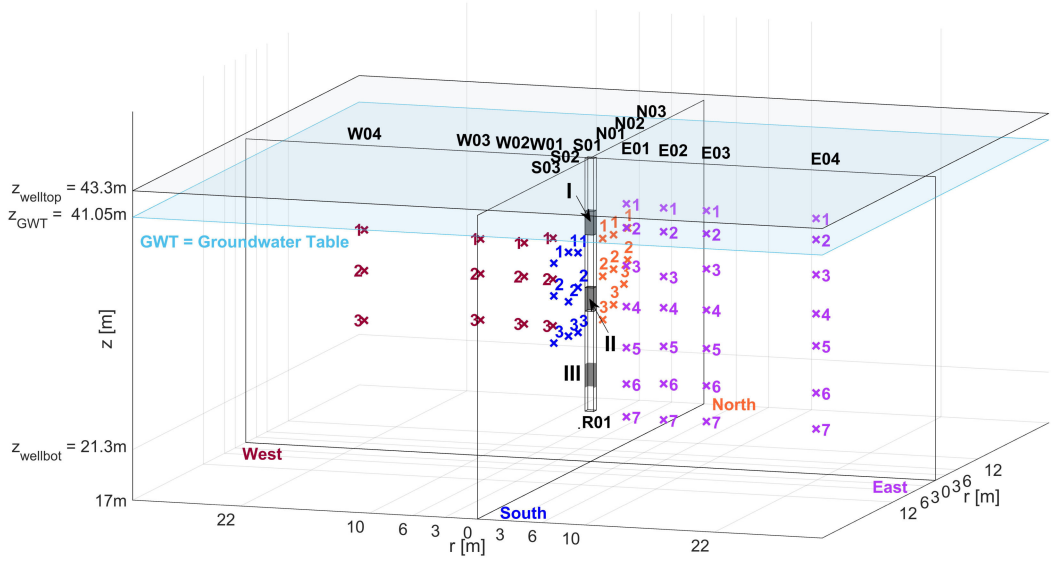

Figure 5.3: Overview of the measurement network installed at the field site. The vertical coordinate is measured from the base of the aquifer, the horizontal coordinates from the pumping well R01 in the middle, illustrated by its well screens.

All observation wells to the east of R01 are continuous multi-channel tubing wells (CMT-wells; Einarson and Cherry, 2002), with seven individual channels with a width of ∼1 cm. The CMT-wells were installed using percussion drilling. Their depths are above, at, and below the elevation of the pumping-well screens I, II, and III. Starting from the top to the bottom, the screen openings are enumerated from 1 to 7 (Figure 5.3). While screen openings 1 to 6 are 0.4 m long, the lowermost screen opening 7 has a length of 0.1 m. Similar to the pumping well, each CMT well was designed with a filter pack consisting of 0.09 m-thick coarse gravel at the depths of the filter-screen sections,

extending 0.2 m above and below each section. The individual filter-pack sections are connected above and below to a bentonite fill.

In total, the monitoring network installed at the field site consists of 58 observation points.

To determine the well efficiency of pumping well R01, a step-drawdown pumping test was performed after completion of the well. In three pumping steps, lasting for 8 h each, water was extracted with a pumping rate of 0.01 m^3/s, 0.02 m^3/s and 0.036 m^3/s, respectively. In each pumping step, water was extracted simultaneously from all three well-screens sections I, II, and III using a frequency-controlled submersible pump (GRUNDFOS). Figure 5.4 shows that the well efficiency decreases with increasing pumping rate.

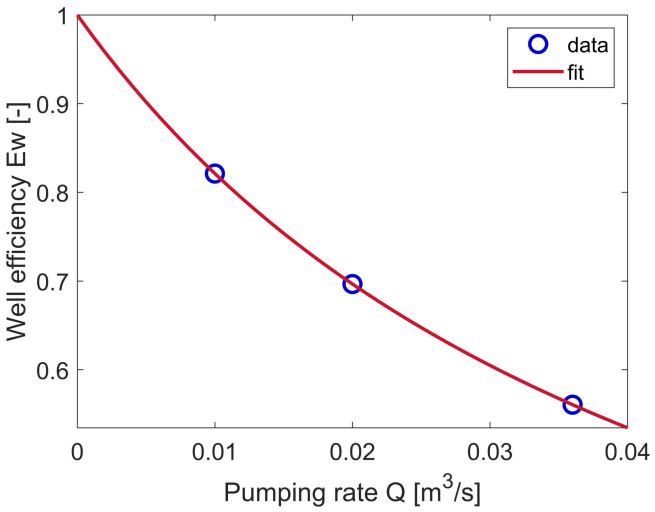

Figure 5.4: Estimation of the well efficiency of pumping well R01 (Stoll, 2020).

5.2 Hydraulic Tests

As a proof-of-concept, three series of short-term pumping tests were performed by extracting water successively from the screen sections I, II, and III of pumping well R01 with a frequency-controlled submersible pump (GRUNDFOS). To prevent water inflow from the adjacent screen sections, a straddle packer system was introduced into the well

and placed above and/or below the active screen section. On account of the large diameter of pumping well R01, the straddle packer was designed as a customized system. The packer system can be adapted in its configuration according to water extraction in the upper, middle or lower well-screen section. Figure 5.5 shows the different compositions of the straddle packer.

For each pumping depth, multiple pumping tests p_t were performed with different

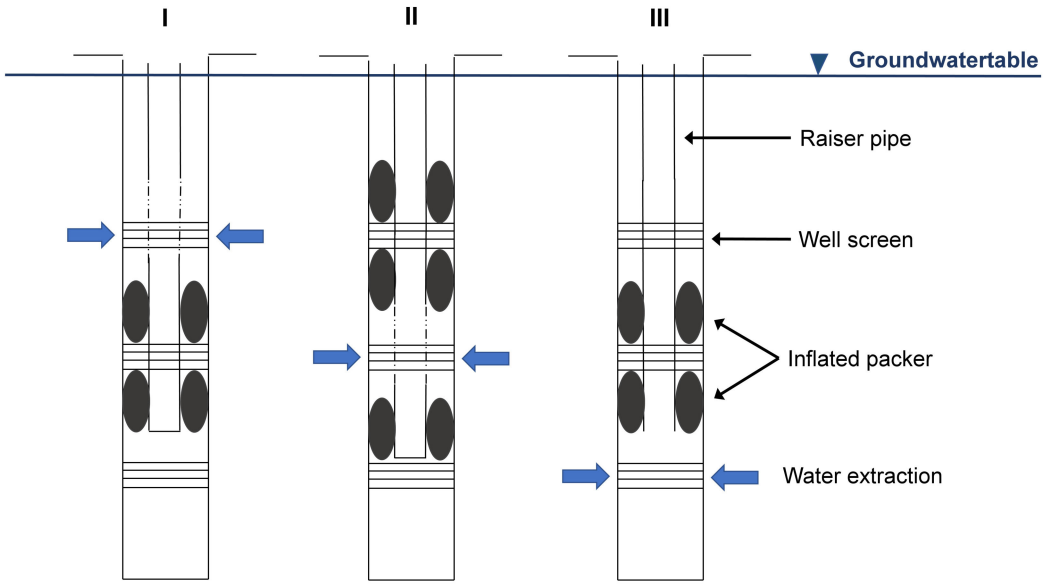

Figure 5.5: Configurations of the straddle packer according to water extraction from well-screen section I, II or III.

pumping rates $Q_w(p_t)$. All experiments belonging to the same extraction interval constitute a specific hydraulic test y. Details on the number of pumping tests belonging to the individual hydraulic tests and the applied pumping rates are listed in Table 5.1.

Table 5.1: Key data of the pumping test series.

	I-Upper screen	II-Middle screen	III-Bottom screen
Hydraulic test y	1	2	3
Number of pumping tests p_t	5	10	7
Reduced number of pumping tests p_r	4	7	4
Time to reach steady-shape behavior t_{trim} **[s]**	1800	1680	1700
Range of $Q_w(p_t)$**[L/s]**	10.0 – 10.5	5.5 – 10.9	17.8 – 19.0

During each pumping test, the transient drawdown response was measured in all available observation points. Fiber-optic pressure transducers (FISO Technologies, Inc.,

Quebec, QC, Canada) were used to record the drawdown response in 14 of the 28 channels available from the CMT wells. I manually measured the drawdown response in the remaining channels using an e-tape. In all 1-inch piezometers of the three-well-bundles the drawdown response was recorded using different types of piezoresistive data loggers. Additionally, the barometric pressure was recorded throughout the complete test period. Detailed information on sensor type, sensor resolution and sampling interval of the different measurement devices is given in Table 5.2.

Table 5.2: Details on the sensors used to measure the drawdown response to water extraction in pumping well R01.

Sensor Type	**Specification**	**Resolution [mm]**	**Sampling Frequency [Hz]**
Fiber-optic pressure transducer	FOP-MPZ-NS-1173B *FISO Technologies*	1	0.2
Absolute pressure transducer	Micro-Diver TD-Diver CTD-Diver *Eijkelkamp Soil & Water*	3	0.2
Relative pressure transducer	Dipper-PT *SEBA Hydrometrie*	-	0.033
Water level meter	Manual measurement	-	-
Barometric pressure transducer	Baro-Diver *Van Essen Instruments*	0.3	0.0033

Due to technical problems with one data logger the total number of active observation points in each hydraulic test was reduced to 57.
In order to monitor the stability of the pumping rate, I regularly took flowmeter readings and observed the drawdown within the pumping well using a pressure transducer. The latter in-well drawdown measurements are not considered in data analysis later on because these measurements are affected by pump-induced pressure variations and well-skin effects.

5.3 Data Processing

The pumping tests resulted in a total of 1334 drawdown curves, from which defective datasets caused by strong instabilities of pressure transducers or the pumping rate

were eliminated. Figure 5.6 shows exemplary datasets of what was defined as defective datasets. In general, datasets showing a drawdown signal that evidently does not correspond to the true signal of pumping were excluded. As shown in Figure 5.6, the instability of pressure transducer recordings can be expressed differently, including a steady decrease of drawdown (lowermost blue and yellow drawdown curves in Figure 5.6), a continuous trend in drawdown (green drawdown curve in Figure 5.6) or no response to pumping at all. With respect to flow rate stability, the entire pumping phase

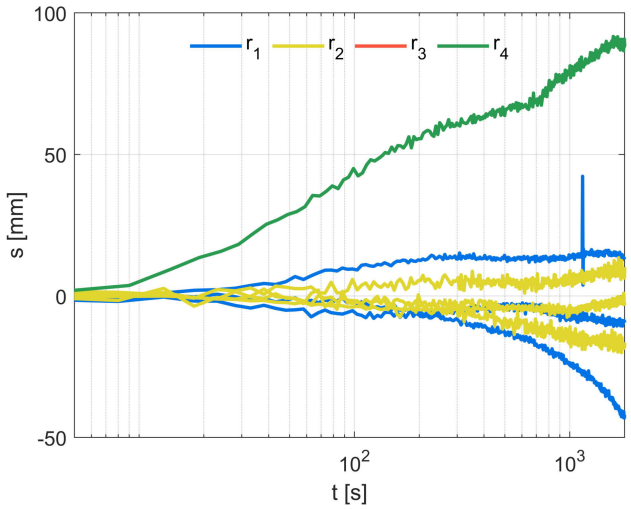

Figure 5.6: Exemplary datasets excluded from model calibration due to instable pressure transducer recordings.

was assessed including the late-term steady discharge. In preceding tests, e.g., the step-drawdown pumping test outlined in section 5.1, and during the hydraulic tests described in section 5.2, it was examined that the system shows an incredibly quick response to changes in the flow rate. Thus, it was possible to identify flow rate instabilities not only directly from flow rate readings but also from observation data.

After the elimination of defective datasets, each curve was corrected to consider barometric pressure variations. Considering that drawdown is linearly proportional to the pumping rate, a scaling was performed for drawdown curves from pumping tests with different extraction rates to those of a harmonized rate Q_h of 0.01 m^3/s. The scaling is based on the mean discharge observed during the pumping phase up to timepoint t_{trim} (Table 5.1). These scaled curves were used to compare the multiple hydraulic responses observed at each observation point and hydraulic test, and to check the data

reproducibility. Pumping tests in which the drawdown curves predominantly failed the reproducibility test, were discarded, reducing the total number of pumping tests from p_t to p_r of each hydraulic test (Table 5.1).

In a next step, I consider the steady-shape pumping regime of the field tests in which the absolute value of drawdown still changes but the hydraulic-head differences between measurement locations remain constant (Bohling et al., 2002, 2007). By analyzing the steady-shape pumping regime, uncertainties can be avoided that are associated with the recorded transient behavior of the hydraulic responses, and the intensive computational requirements of simulating transient groundwater flow is reduced.

In the reduced set of pumping tests p_r, the steady-shape regime was defined by identifying the time (t_{trim}) at which changes in drawdown differences between observation locations could be neglected (see Table 5.1). Figure 5.7 shows an exemplary dataset to demonstrate the identification of the steady shape pumping regime. Figure 5.7a shows the absolute drawdowns versus pumping time until timepoint t_{trim} = 1800 s obtained with pumping test p_t = 5 of hydraulic test I, i.e. when water was extracted from the upper screen in pumping well R01. The drawdowns increase with time and decrease with increasing radial distance to the well (different colors in Figure 5.7a). Figure 5.7b shows the difference in drawdown between all observation points and observation point W04.3, chosen as an exemplary reference point. Contrary to the absolute drawdowns, the drawdown differences appear to remain constant after a pumping time of ∼15 min, that is, prior to the selected timepoint t_{trim}.

Additionally, I computed the logarithmic derivative of drawdown as a function of time in log-log scale (Renard et al., 2009) using the differentiation algorithm after Bourdet et al. (1989) to assess the change in deviation at each observation point (Figure 5.7c). Figure 5.7c shows that for p_t = 5 of hydraulic test I the logarithmic derivatives stabilize prior to timepoint t_{trim} with deviations smaller than 1 mm. The stabilization of logarithmic derivatives indicates that most likely the infinite acting radial flow phase was reached (Renard et al., 2009). With this, the aquifer can be assumed of infinite areal extent at timepoint t_{trim} with no influences of well bore effects or external boundaries on drawdown.

In a last step, I averaged the drawdown s_{obs} at timepoint t_{trim} from all pumping tests p_r available in hydraulic test y. By that, a single averaged drawdown measurement

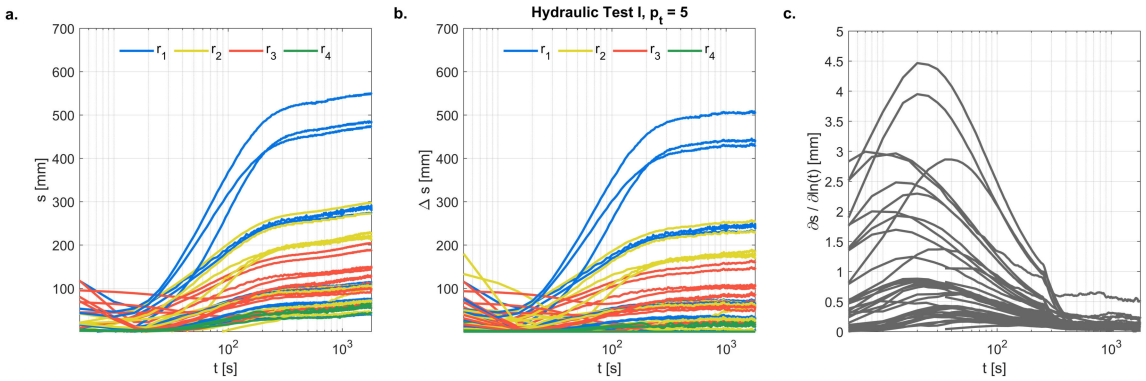

Figure 5.7: Dataset of pumping test $p_t = 5$ when water was extracted from the upper screen (hydraulic test I) showing a. the absolute drawdown measurements versus pumping time. b. the difference in drawdown measured in the different observation wells and reference well W04.3. c. the logarithmic time derivative of drawdown.

s_{meas} was obtained for each observation point k in each hydraulic test y

$$s_{meas}(y,k) = \frac{1}{p_r}\sum_{i=1}^{p_r} s_{obs,i}(y,k) \tag{5.1}$$

and an associated standard deviation σ_{repr} as metric of the reproducibility of each measurement:

$$\sigma_{repr}(y,k) = \sqrt{\frac{1}{p_r - 1}\sum_{i=1}^{p_r}(s_{obs,i}(y,k) - s_{meas}(y,k))^2} \tag{5.2}$$

I denote σ_{repr} the reproducibility error, which contributes to the overall error of the measurements but does not include systematic errors in the data (e.g., due to the misplacement of observation points) or in the conceptual model (e.g., due to disregarding horizontal heterogeneity).

To quantify the variability among the measurement points, I clustered observation points which are arranged in different directions to the pumping well, but coincide by ±0.6 m and ±0.5 m in their r- and z-coordinates, respectively, and compare the associated drawdown measurements s_{meas}. These cluster ranges are considered to be realistic since intended measurement locations of observation points may be misplaced in the installation of observation wells (Maier et al., 2020). I computed the mean value

μ_{c_p} and standard deviation σ_{c_p} from all measurements n_{c_p} available in each cluster c_p

$$\mu_{c_p}(y) = \frac{1}{n_{c_p}(c_p)} \sum_{i=1}^{n_{c_p}(c_p)} s_{meas,i}(y, c_p) \tag{5.3}$$

$$\sigma_{c_p}(y) = \sqrt{\frac{1}{n_{c_p}(c_p)} \sum_{i=1}^{n_{c_p}(c_p)} (s_{meas,i}(y, c_p) - \mu_{c_p}(y))^2} \tag{5.4}$$

5.4 Processed Data Set

The previously presented data-processing strategy facilitated the reduction of over 1000 recorded drawdown curves to a set of $n_{meas} = 3 \times 57 = 171$ steady-shape drawdown observations, given that in each of the three hydraulic tests the recording in one observation point failed. Figure 5.8 shows the drawdown measurements s_{meas} rescaled to reflect a harmonized pumping rate of 0.01 m^3/s belonging to the hydraulic test performed at the top (Figure 5.8 a), the middle (Figure 5.8b) and the bottom (Figure 5.8c) screen of the pumping well. The measurements are displayed as a function of the radial distance to the pumping well (different colors in Figure 5.8a-5.8c), while for each radial distance the observations are aligned with elevation (bar placement along the z-axis in Figure 5.8a-c). As expected, Figure 5.8a-Figure 5.8c show that in each hydraulic test the observed drawdown decreases with increasing radial distance to the pumping well and are higher at elevations close to the pumped screen interval. The drawdown observations s_{meas} range between millimeters and meters. The signal strength differs between the three tests even after correcting for different pumping rates. When extracting water from the lower screen (Figure 5.8c), the drawdown does not reach the high values observed in the other tests, whereas the strongest responses result from the hydraulic test with water extraction from the middle screen (Figure 5.8b). These differences may be caused by vertical variations of hydraulic conductivity. In particular, a lower-conductivity layer at a depth close to that of the middle screen could explain higher drawdown values when water is extracted from this screen.

Figure 5.9a shows the rescaled drawdown measurements s_{meas} together with the errors obtained during the reproducibility test σ_{repr} (see section 5.3). Again, Figure 6a

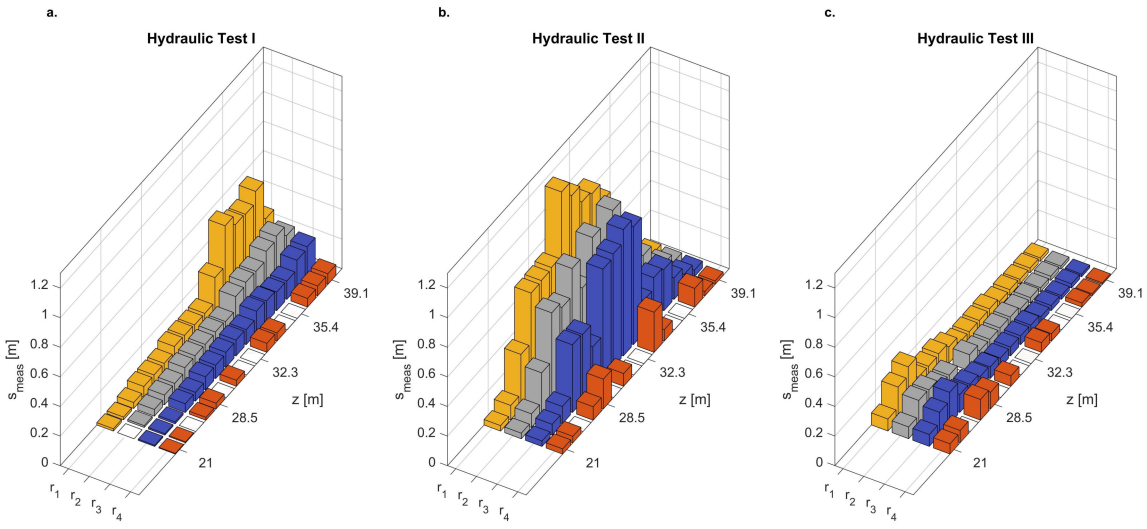

Figure 5.8: Field measurements of the hydraulic tests performed at (a) the top, (b) the middle and (c) the bottom screen of the pumping well. The yellow, grey, blue and orange colors correspond to the radial distances $r_1 \in$[3.3,3.6], $r_2 \in$[6.2,6.8], $r_3 \in$[10.2,10.7] and $r_4 \in$[20.9,21], respectively.

shows that the measurement signal varies between the three hydraulic tests (different colors in Figure 5.9a) while the associated errors are within a similar range for the three tests. In general, the reproducibility errors do not exceed values of 2.2 cm and are smallest for the hydraulic test with water extraction in the lower screen.

A comparison of the errors between the four directions (north, east, south, and west of the pumping well, see different marker symbols in Figure 5.9a) reveals no significant spatial pattern with respect to reproducibility.

Figure 5.9b shows the errors associated with measurements obtained at similar radial distance and depth but in different directions to the pumping well. In contrast to the comparably small reproducibility errors σ_{repr}, the errors associated with horizontal heterogeneity are higher and reach values of up to 10 cm.

5.5 Numerical Model Setup

To simulate the hydraulic tests described in section 5.2, two different groundwater flow models were built. The first model considers a single, homogeneous layer (1-layer

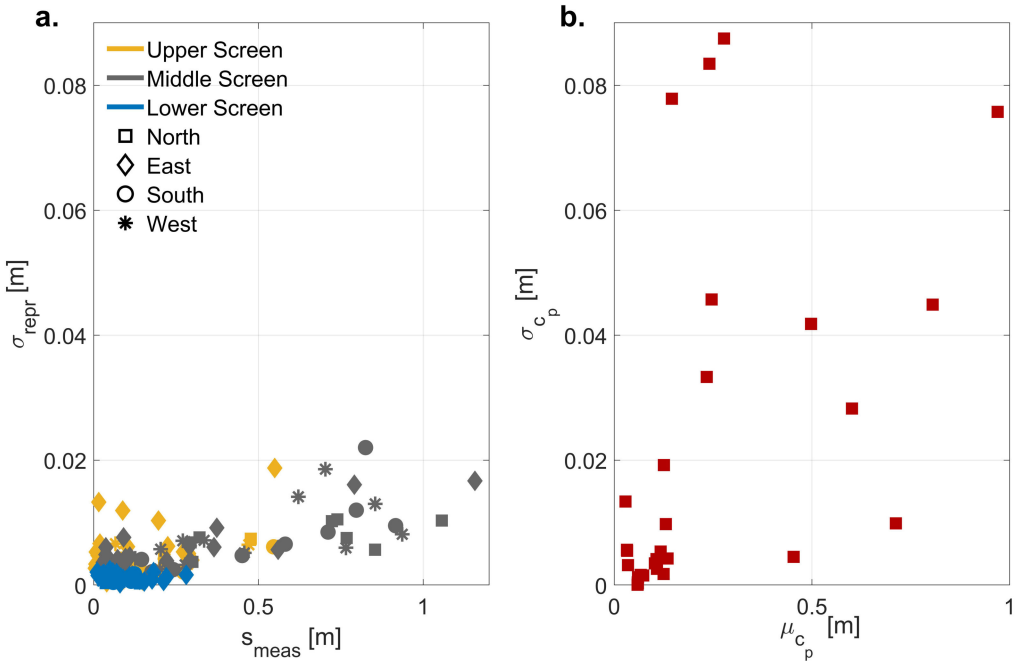

Figure 5.9: a. Measurement errors σ_{repr} resulting from the reproducibility analysis as function of the drawdown measurements s_{meas}. Colors indicate extraction from the different pumping intervals in different depths, marker styles indicate horizontal orientation of the measurement points. b. Measurement errors σ_{c_p} resulting from measurements at similar radial distances and depths but different directions to the pumping well.

model) whereas the second one contains five horizontal layers (5-layer model). Both models account for hydraulic anisotropy in each layer, that is, to each layer a radial and vertical hydraulic conductivity K_r and K_z are assigned, respectively. In both models, the aquifer is treated as confined, which is a justifiable assumption due to the short test durations, and the small drawdowns in comparison to the total thickness of the aquifer.

Figure 5.10 shows the basic model setup. In both models, two separate units are defined that represent the gravel pack (red units in Figure 5.10) and clay fill (gray units in Figure 5.10) of the filter pack installed around the pumping well. To each unit an isotropic hydraulic conductivity K_{gp} and K_{cl} are assigned, respectively. The black dashed horizontal lines in Figure 5.10 indicate the layer boundaries considered in the 5-layer model. I used the existing field description of the drill core of pumping well R01 (Figure 5.10, Figure 5.11a) to define the actual location of the layers 1 to 5 in the 5-layer model, numbered from the aquifer top to the bottom. As a selection criterion, I assumed

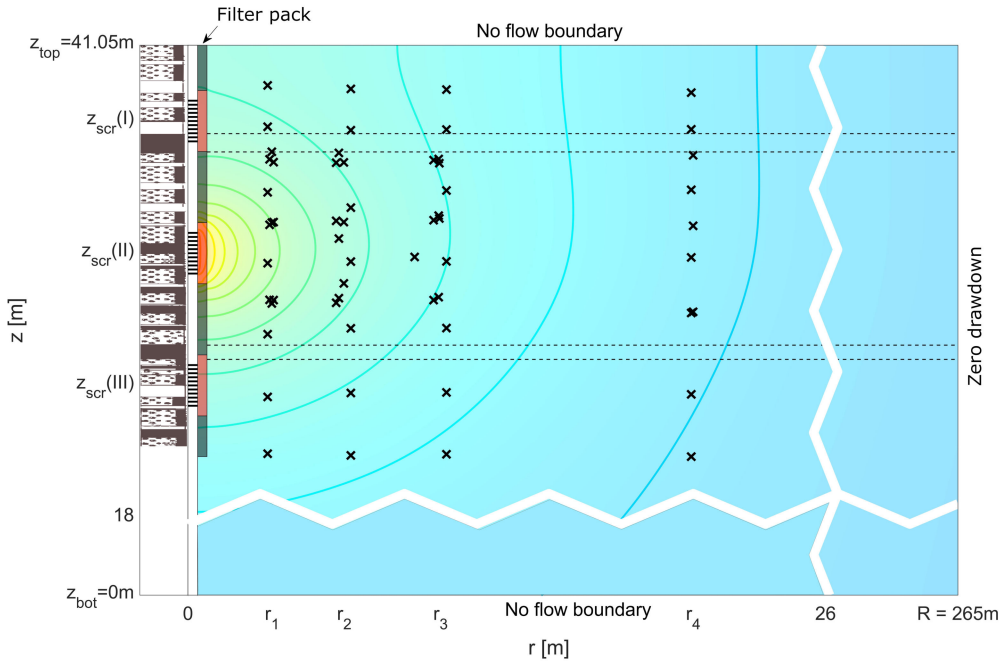

Figure 5.10: Conceptual representation of the 1-layer and 5-layer model, distinguished by dashed horizontal lines. The drilling profile is displayed left to the pumping well showing lithologic information at the pumping well location. Also, the gravel pack (red) and clay seal (gray) are displayed right to the pumping well. The black crosses correspond to individual observation points, placed at radial distances $r_1 \approx 3.5$ m, $r_2 \approx 6.5$ m, $r_3 \approx 10.5$ m and $r_4 \approx 21$ m. The model domain reaches from the saturated aquifer thickness at $z_{top} = 41.05$ m to the aquifer bottom at $z_{bot} = 0$ m. Note that axes are not to scale and that the contour lines of drawdown correspond to an exemplary response in a homogeneous aquifer system, i.e. considering a 1-layer model.

a continuous horizontal layer to be present at depths in which the sand fraction clearly exceeds that of the gravel and spans more than half a meter in thickness (Figure 5.11b). I defined the threshold criterion of 0.5 m assuming that a layer of larger thickness is more likely to occur more or less consistent with depth throughout the domain than a layer of smaller thickness and thus, can be resolved in a model in the first place.

To simulate the hydraulic tests, I used a Finite-Element model implemented in Matlab that solves the axisymmetric steady-state groundwater flow equation on rectangular elements. The radial grid spacing increases logarithmically, whereas the vertical resolution is uniform with a grid spacing of 0.17 m. To obtain simulation results at observation points that did not fall onto nodes of the grid, a bilinear interpolation was performed consistent with the Finite Element formulation.

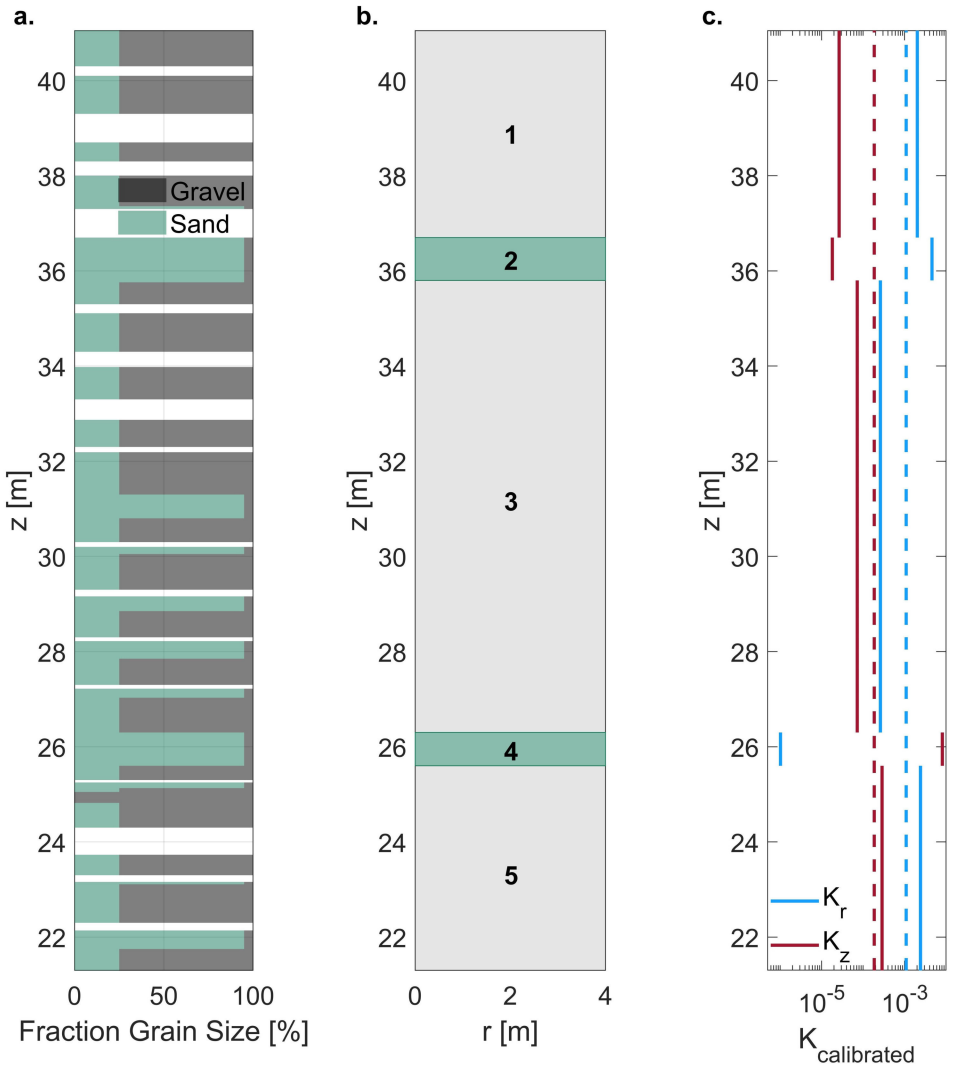

Figure 5.11: a. Relative gravel and sand fractions from the field description of the drilling core of the large diameter well R01 at the field site Kappel-Grafenhausen (white gaps = core loss). b. Definition of five horizontal layers considering depths of more than half a meter in thickness in which the sand fraction clearly exceeds the gravel fraction. c. Calibrated radial (blue) and vertical hydraulic conductivities (red) of the 1-layer (dashed lines) and 5-layer model (solid lines).

5.6 Model Calibration

After setting up the groundwater flow models, the 1-layer and 5-layer models were calibrated independently using the Trust-Region Reflective Least-Square method of the

function `lsqnonlin` in the optimization toolbox of Matlab (Coleman and Li, 1996). All three hydraulic tests were considered jointly in the calibration, with $n_{meas} = 171$ drawdown observations.

As mentioned in section 2.2 and 5.3, I consider a steady-shape pumping regime, in which drawdown differences between observation locations remain constant. Typically, this requires the specification of pairs of observation points by either setting one observation location as the superordinate reference point (Maier et al., 2020) or by considering all feasible pairs of observation points (Bohling et al., 2002). Each field measurement, however, is subject to measurement errors of different types, including measurement noise or the misplacement of observation wells (Maier et al., 2020). I have tested the effect of considering different observation points as the reference point, yielding different model-calibration results due to measurement error. To avoid the propagation of uncertainties in the generation of pairs of observation points, I included a virtual reference point in the model calibration. That is, for each hydraulic test the simulated drawdown was considered as s_{sim}=$|s_t - s_{ref}|$ in which s_t is the simulated steady-state drawdown and s_{ref} is the drawdown at a virtual reference point, that is also estimated.

Then, the differences between the simulated and measured drawdowns s_{sim} and s_{meas} were computed and normalized by the error σ_i of each measurement i, which is defined by an error model discussed below.

The objective function φ was to minimize the sum of squared normalized residuals:

$$\varphi = \sum_{i=1}^{n_{meas}} \left(\frac{s_{sim,i}(\mathbf{p}) - s_{meas,i})}{\sigma_i} \right)^2 \tag{5.5}$$

in which $\mathbf{p}$ is the parameter vector including the logarithms of K_r and K_z of all horizontal layers considered, the log-hydraulic conductivities K_{gp} and K_{cl} of the filter pack, and the drawdown s_{ref} for each of the three hydraulic tests. Thus, in total, the 1-layer and 5-layer models included $n_{par} = 7$ and $n_{par} = 15$ calibration parameters, respectively.

To normalize the residuals of simulated and measured drawdowns in model calibration, I chose an error model that accounts for the combined effects of the reproducibility error, a potential measurement bias (e.g., due to the misplacement of the observation points), and most importantly the model-conceptual error (e.g., due to a suboptimal

definition of layers or lacking 3-D heterogeneity). In essence, I don't claim to have defined models that are perfect representations of reality and accept misfits that are bigger than the error of the measurements themselves for the sake of keeping the hydrogeological models comparably simple and the fitted parameters meaningful. In this framework, a heteroscedastic error model must be defined with a set of parameters that become part of the fitting procedure. As different models have different deficiencies, they have different model errors, and judging the quality of the different models is based on the fitted coefficients of the error model.

I have tested different error models and upon analyzing the residuals, I have chosen the following parameterization:

$$\sigma = a + \frac{b \cdot s_{meas}^2}{c + s_{meas}} \tag{5.6}$$

with a, b and c being the error-model parameters. This specific error model starts off with a constant error with $\lim_{s_{meas}\to 0} \sigma = a$ corresponding to the absolute error, then shows a quadratic increase with the measurement, and converges to a linear dependence on s_{meas} for large values with $\lim_{s_{meas}\to \infty} \frac{\sigma}{s_{meas}} = b$ corresponding to the relative error. The parameter c quantifies how quickly the error model converges from the measurement-independent to the linear regime.

The error model parameters were determined by calibrating the 1-layer and 5-layer model according to the expectation-maximization method (Dempster et al., 1977). That is, the objective function was iteratively minimized with the Trust-Region Reflective Least-Square method of the function `lsqnonlin` in the optimization toolbox of Matlab (Coleman and Li, 1996) and the error-model parameters were updated by performing a least-square fit of the error model to the absolute residuals $|s_{sim}(\mathbf{p}) - s_{meas}\ |$ of the model fit to the measured drawdown s_{meas}. With this, the error model parameters a, b and c were included in the optimization process along with all model parameters. The iterative calibration procedure was considered completed when the change in all model and error parameters was less than 1%.

The comparison of the different models is now not based on meeting the observations within the measurement error but on the magnitude of the model error needed to accept the different models.

After fitting the models, the associated standard deviation $\hat{\sigma}_{p_i}$ of estimation of the model parameter i was computed by linearized error propagation:

$$\hat{\sigma}_{p_i} = \sqrt{\mathbf{C_{pp}}(i,i)} \tag{5.7}$$

with the parameter covariance matrix $\mathbf{C_{pp}}$ computed by:

$$\mathbf{C_{pp}} = (\mathbf{J}^T \mathbf{C_{yy}}^{-1} \mathbf{J})^{-1} \tag{5.8}$$

in which the Jacobian $\mathbf{J}$ contains the partial derivatives of all simulated measurements with respect to all parameters, and $\mathbf{C_{yy}}$ is the diagonal matrix of the squared errors according to the error model. Because the parameters a and b of the error model are bigger if the model shows larger misfits, the resulting parameter standard deviations of estimation are also bigger.

5.7 Results and Discussion

5.7.1 Goodness of Model Calibration

Figure 5.12a and Figure 5.12b show the absolute differences between simulated and measured drawdowns $|s_{sim} - s_{meas}|$ versus the measured drawdowns s_{meas} of the 1-layer and 5-layer model calibration considering the final error model update. Comparing Figure 5.12a and Figure 5.12b shows that the differences between simulated and measured drawdowns are significantly higher for the 1-layer than for the 5-layer model. In both cases, the error-model fit according to eq. (5.6) is able to capture the majority of measurements and its errors, while some outliers exist. Figure 5.12c and Figure 5.12d present a comparison between measured and simulated drawdown values, s_{meas} and s_{sim}, for the best fitting 1-layer and 5-layer models associated with the error models determined in Figure 5.12a and Figure 5.12b, respectively. Figure 5.12c reveals that the 1-layer model systematically underestimates the drawdown in the second hydraulic test, in which water was extracted from the middle screen section, whereas the 5-layer model can decently fit all three hydraulic tests (Figure 5.12d). As discussed in section 5.4, the drawdown values were higher in the second test series than in the tests where water was extracted from the bottom and top screen, respectively. The 5-layer model can

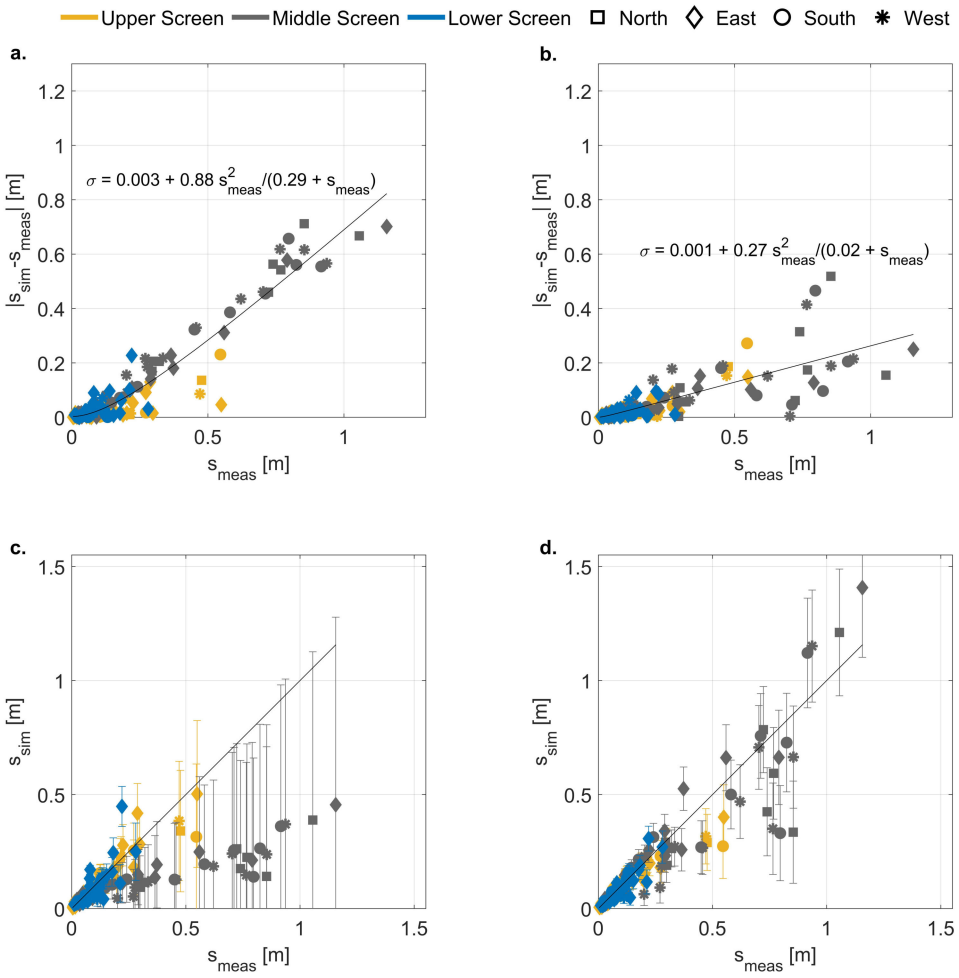

Figure 5.12: Assessment of model results: Absolute difference between simulated and measured drawdown versus the measured drawdowns of the 1-layer and 5-layer model and the thereto fitted error models (a and b). Field measurements versus simulated results of the 1- and 5-layer models (c and d) with errorbars according to the error model. The black diagonal lines represent the 1:1 identity lines.

reproduce this pattern by fitting a lower horizontal hydraulic-conductivity value to the middle depth of the aquifer (see Table 5.3), whereas the 1-layer can either fit the high drawdown values of the second hydraulic test or the smaller drawdown measurements of the first and third tests.

In general, the 5-layer model meets the majority of measured drawdowns (Figure 5.12d), with only a few measurements falling far off the identity line. As indicated by the different marker styles in Figure 5.12d, there is no clear relationship between the

directions of the measurement locations (north, east, west, and south of the pumping well) and the tendency towards over- or underestimating the measured drawdown values, obviating horizontal anisotropy.

5.7.2 Fitted Parameter Values

Table 5.3 and Table 5.4 list the parameters estimated for both models. As mentioned in section 5.6, the model parameters included the radial and vertical hydraulic conductivities K_r and K_z of each horizontal layer considered in the model, the hydraulic conductivities of the gravel fill K_{gp} and clay fill K_{cl} of the filter pack, and the drawdown s_{ref} at a virtual reference point to avoid computing the drawdown differences between true observation points in the steady-shape regime. While Table 5.4 contains the fitted values of s_{ref}, they don't have any real physical meaning.

Figure 5.11c includes the radial (blue lines) and vertical (red lines) hydraulic conductivities estimated for the 1-layer model (dashed lines) and the 5-layer model (solid lines). In both cases, the radial hydraulic conductivities are higher than the vertical counterparts, except for layer 4 of the 5-layer model (Table 5.3).

The radial and vertical hydraulic log-conductivities estimated for the sand layer 4 have considerably higher associated standard deviations $\hat{\sigma}(lnK)$ of estimation than all other conductivity estimates (Table 5.3), indicating that the measurements are insensitive to the conductivities estimated for that layer. Also, the fitted conductivity values of the 5-layer model reveal that the sand layer 2 is quite similar to layer 1. This implies that the sand layers, which were delineated by the analysis of the drilling core, may not necessarily constitute distinct individual layers, but rather represent transition zones between the three main aquifer segments. In hindsight, the investigated aquifer portion may have been subdivided into three main sections, without much impact. The horizontal hydraulic conductivity K_r for the top and bottom of the investigated aquifer portion show significant higher values than the middle section, whereas the fitted vertical conductivity K_z systematically increases with depth. The reduced horizontal conductivity of the middle section (layer 3) can explain the larger drawdown values in the second test series, in which groundwater was extracted from the middle screen.

Table 5.3: Calibrated radial and vertical hydraulic conductivities and the associated standard deviations $\hat{\sigma}$ of estimation of each horizontal layer in the 1-layer and 5-layer models.

	Layer	K_r**[m/s]**	$\hat{\sigma}_{\mathbf{lnK_r}}$	K_z**[m/s]**	$\hat{\sigma}_{\mathbf{lnK_z}}$
1-layer model		1.1×10^{-3}	0.0581	1.9×10^{-4}	0.154
5-layer model	**1**	2.0×10^{-3}	0.581	2.6×10^{-5}	1.063
	2	4.6×10^{-3}	0.912	1.8×10^{-5}	0.605
	3	2.6×10^{-4}	0.171	7.2×10^{-5}	0.190
	4	1.0×10^{-6}	195.942	8.1×10^{-3}	37.113
	5	2.4×10^{-3}	0.164	2.9×10^{-4}	0.3499

In both models, the estimated hydraulic conductivity of the gravel pack K_{gp} is higher than any conductivity value of the aquifer, whereas the fitted conductivity of the clay seal K_{cl} is so small that it can be considered impermeable (Table 5.4).

Table 5.4: Calibration results of the locally anisotropic 1-layer and 5-layer models.

	1-layer model	**5-layer model**
$\mathbf{K_r^{eff}}$**[m/s]**	1.1×10^{-3}	1.9×10^{-3}
$\mathbf{K_z^{eff}}$**[m/s]**	1.9×10^{-4}	9.5×10^{-5}
$\vartheta[-]$	5.9	19.7
$\mathbf{K_{gp}}$**[m/s]**	1.5×10^{-1}	8.8×10^{-1}
$\mathbf{K_{cl}}$**[m/s]**	2.2×10^{-8}	7.9×10^{-8}
$\mathbf{s_{ref,I}/s_{ref,II}/s_{ref,III}}$	0.078 / 0.058 / 0.069	0.035 / 0.029 / 0.026
a[m]	0.003	0.001
b$[-]$	0.88	0.27
c[m]	0.29	0.02

Finally, I compare the determined absolute and relative errors of the 1-layer and 5-layer models (Table 5.4) to assess the goodness of the model-calibration fit. While the 1-layer model shows an absolute error of $a = 3$ mm and a relative error of $b = 88\%$, the 5-layer model has an absolute error of 1 mm and a considerably lower relative error of 27%, proving to be the preferred model choice.

5.7.3 Assessing the Effective Conductivity Estimates

Subsequent to model calibration, the radial and vertical conductivity estimates of all available horizontal layers were fully upscaled to obtain the effective radial and effective

vertical hydraulic conductivities K_r^{eff} and K_z^{eff}, respectively (Table 5.4). In the 1-layer model, the fitted horizontal and vertical conductivities K_r and K_z correspond to the effective values K_r^{eff} and K_z^{eff}, respectively. The estimated effective horizontal conductivities K_r^{eff} of the 5-layer model is about twice as high as the estimated value of the 1-layer model (see Table 5.4), while the upscaled vertical conductivity K_z^{eff} of the 5-layer model is about half the fitted value of the 1-layer model. As a consequence, the anisotropy ratio $\vartheta = K_r^{eff}/K_z^{eff}$ differs between the two models by a factor of more than 3 (≈ 6 versus ≈ 20). Both estimated anisotropy ratios are within reasonable ranges expected for fluvial deposits (Freeze and Cherry, 1979; Kruseman and de Ridder, 1994).

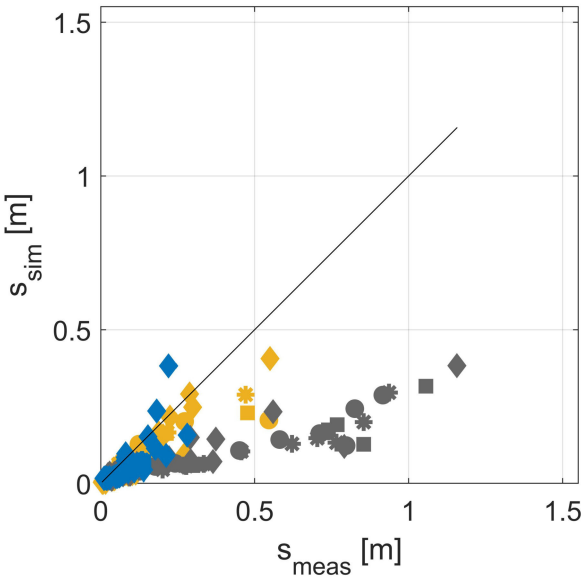

Figure 5.13: Simulated drawdowns versus measured drawdowns when running the 1-layer model with the effective conductivity estimates obtained from the 5-layer model calibration. The black diagonal line present the 1:1 identity line.

To examine the value of the effective conductivity estimates, I have run the 1-layer model considering the effective conductivity results obtained by fully upscaling the 5-layer model. Figure 5.13 shows the resulting simulated drawdowns versus the measured drawdowns and the 1:1 identity line (black line in Figure 5.13). I compare the measurement fit of the 1-layer model when using the effective values from the 5-layer model (Figure 5.13) with the measurement fit based on the effective values resulting from the 1-layer model calibration (Figure 5.12c). Comparing Figure 5.13 and Figure 5.12c indicates that considering a homogeneous model with the effective conductivity estimates

obtained from a multilayer model does not yield any improvement nor any deterioration in fitting the true drawdown measurements. This suggests that the effective horizontal hydraulic conductivity and effective vertical hydraulic conductivity are little informative when neglecting vertical differences of the hydraulic conductivity within the aquifer.

5.7.4 Comparison to Locally Isotropic Models

To check whether the horizontal layers should really be locally anisotropic, the 1-layer and 5-layer models were calibrated considering the hydraulic conductivity of each available layer to be isotropic, i.e. $K_r = K_z$. By this, the number of model parameters is reduced to 6 and 10 for the 1-layer and 5-layer model, respectively. For differentiation, I now refer to the groundwater flow models as either locally isotropic or anisotropic 1-layer and 5-layer models. Figure 5.14a and b show the absolute differences between the simulated and measured drawdowns of the locally isotropic 1-layer and 5-layer models considering the final error model update. Similar to the results of the locally anisotropic model calibrations, the locally isotropic 5-layer model reveals smaller differences between simulated and measured drawdowns than the 1-layer model counterpart.

Figure 5.14c and d show the simulated versus the measured drawdowns for the calibration result of the locally isotropic 1-layer and 5-layer models considering the error-model fit displayed in Figure 5.14a and b. The systematic underestimation of the large-drawdown values by the locally anisotropic 1-layer model (see Figure 5.12c) is enhanced when assuming a single isotropic conductivity value of the entire formation (see Figure 5.14c). This is consistent with a relative error of 100% for the isotropic 1-layer model listed in Table 5.5. Quite obviously, assuming a uniform isotropic conductivity is not appropriate.

Figure 5.14d shows the fit of the locally isotropic 5-layer model, which performs better than both the isotropic and anisotropic 1-layer models. The comparisons between Figure 5.12d and Figure 5.14d as well as between the coefficients of the error model (27% relative error for the locally anisotropic model versus 37% for the locally isotropic counterpart, in both cases with small to negligible absolute errors) indicates that accounting for different horizontal and vertical hydraulic conductivities within the layers improves fitting the depth oriented drawdown observations.

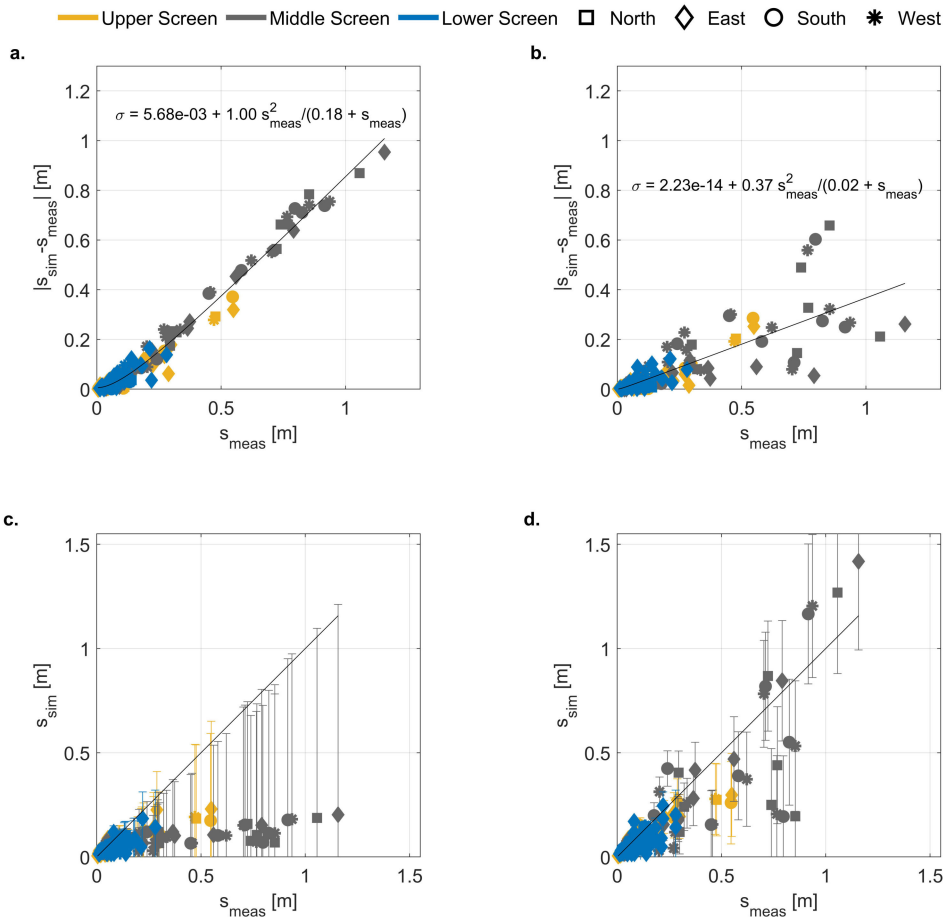

Figure 5.14: Assessment of model results when considering isotropic hydraulic conductivity: Absolute difference between simulated and measured drawdown versus the measured drawdowns of the 1-layer and 5-layer model and the thereto fitted error models (a and b). Field measurements versus simulated results of the 1- and 5-layer models (c and d) with errorbars according to the error model. The black diagonal lines represent the 1:1 identity lines.

Table 5.5 also lists the conductivity estimates and the associated standard deviations of estimation for each layer considered in the locally isotropic 1-layer and 5-layer models. The conductivity estimates of the individual layers of the locally isotropic 5-layer model of layers 2 and 3 are very similar, putting the choice of the layers into question. As discussed in section 5.7.2, layers 2 and 4 may not necessarily constitute individual horizontal layers throughout the aquifer but rather represent passages between the adjoining layers. Comparing the conductivity estimates of the locally isotropic and

Table 5.5: Calibrated isotropic hydraulic conductivities and the associated standard deviations $\hat{\sigma}_{lnK}$ of estimation of each horizontal layer in the 1-layer and 5-layer models. Error model parameters of the isotropic 1-layer and isotropic 5-layer model.

	layer	**K[m/s]**	$\hat{\sigma}_{\mathbf{lnK}}$	**a[m]**	**b[−]**	**c[m]**
1-layer model		1.4×10^{-3}	0.142	0.006	1.00	0.18
5-layer model	**1**	2.9×10^{-3}	0.132			
	2	1.3×10^{-4}	0.320			
	3	1.3×10^{-4}	0.109	2.23×10^{-14}	0.37	0.02
	4	9.2×10^{-3}	0.642			
	5	1.0×10^{-3}	0.425			
		$\mathbf{K_r^{eff}}$**[m/s]**		1.10×10^{-3}		
		$\mathbf{K_z^{eff}}$**[m/s]**		3.72×10^{-4}		

anisotropic 5-layer models shows that the fitted isotropic conductivities of layers 1, 3 and 5 are similar to the radial conductivities of the locally anisotropic model with similar standard deviations.

Upon full upscaling, the effective hydraulic conductivity of the locally isotropic 1-layer model is with 1.4×10^{-3} m/s in a similar range as the effective radial conductivity resulting from the locally anisotropic 1-layer model calibration (for comparison see Table 5.4). In case of the locally isotropic 5-layer model the directional dependence of the effective hydraulic conductivity K_r^{eff} is with 1.1×10^{-3} m/s similar to the anisotropic counterpart but K_z^{eff} is with 3.72×10^{-4} m/s four times larger (for comparison see Table 5.4). Comparing the resulting anisotropy ratios among the locally anisotropic 5-layer model ($\vartheta = 19.7$) and the locally isotropic 5-layer model ($\vartheta = 3.0$) indicates that neglecting local differences in radial and vertical hydraulic conductivity significantly reduces the anisotropy ratio on larger scales.

For both the locally isotropic and anisotropic 5-layer models the fitted error models approximate a linear error model with an intercept of zero. For consistency I have applied the same error model to the 1-layer and 5-layer models, yet, the significance of the absolute error may be questioned for the 5-layer models.

5.7.5 Updating the Model Setup

As discussed in section 5.7.2 and section 5.7.4, the sand layers 2 and 4 defined in the locally anisotropic and locally isotropic 5-layer models may be merged to the three main

aquifer segments rather than considering them as individual horizontal layers. To test this hypothesis, I setup a groundwater flow model of three horizontal layers (3-layer model) by removing the lower and upper boundaries of sand layer 2 and 4, respectively. Similar to the previously considered locally anisotropic groundwater flow models, the 3-layer model was calibrated estimating the radial and vertical hydraulic conductivities K_r and K_z of each horizontal layer, the conductivities of the gravel fill and clay fill, K_{gp} and K_{cl}, and the drawdown at a virtual reference point s_{ref}. Again, an iterative calibration procedure was performed in which the error model parameters a, b and c were determined. In the following, I compare the results of the 3-layer model to the results of the 5-layer model, both considering local anisotropy.

Figure 5.15a shows the absolute difference between the simulated and measured drawdown values $|s_{sim} - s_{meas}|$ versus the measured drawdown values s_{meas} considering the final error model update. Similar to the 5-layer model, the error model fit of the 3-layer model meets the majority of measurements and its errors with few exceptions (for comparison see Figure 5.12b). Figure 5.15b presents the measured drawdown s_{meas}

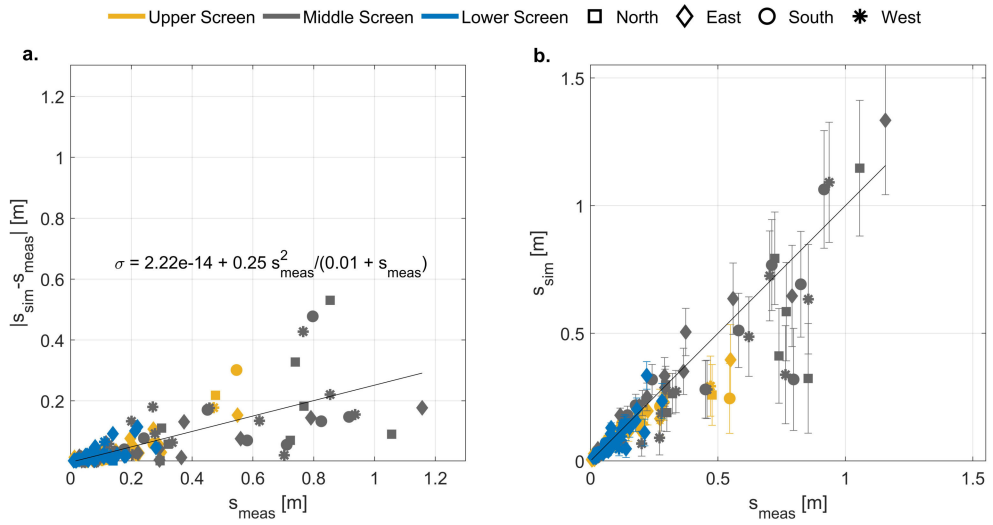

Figure 5.15: Absolute difference between simulated and measured drawdown versus the measured drawdown of the 3-layer model and the thereto fitted error model (a). Field measurements versus simulated results of the 3-layer model (b) with errorbars according to the error model. The black diagonal line represents the 1:1 identity line.

versus the simulated drawdown s_{sim} resulting from model calibration with the 3-layer model considering the error model fit in Figure 5.15a. Comparing Figure 5.15b and 5.12d reveals that both the 5-layer model and the 3-layer model are able to reproduce the true drawdown measurements similarly well, taking into account individual outliers. This indicates that excluding the sand layers 2 and 4 in model calibration does not genuinely affect the reproduction of measured drawdown.

Table 5.6 and 5.7 list the parameters estimated for the 3-layer model. Both the radial conductivities K_r and vertical conductivities K_z estimated for layer 1, 2 and 3 of the 3-layer model (Table 5.6) resemble the respective conductivity estimates of the equivalent aquifer segments assigned to layers 1, 3 and 5 of the 5-layer model at standard deviations of the estimation in the same order of magnitude (for comparison see Table 5.3).

Comparing the effective radial and vertical conductivities K_r^{eff} and K_z^{eff} after full upscaling among the 5-layer and 3-layer model indicates very similar results. The anisotropy ratios predicted with the 3-layer model and 5-layer model are within the same order of magnitude, yet, the predicted anisotropy ratio for the 3-layer model is with $\vartheta = 15.9$ smaller than the predicted anisotropy ratio of $\vartheta = 19.7$ of the 5-layer model.

Table 5.6: Calibrated radial and vertical hydraulic conductivities and the associated standard deviations $\hat{\sigma}$ of estimation of each horizontal layer in the 3-layer model.

	Layer	K_r**[m/s]**	$\hat{\sigma}_{\mathbf{lnK_r}}$	K_z**[m/s]**	$\hat{\sigma}_{\mathbf{lnK_z}}$
3-layer model	**1**	2.6×10^{-3}	0.130	3.2×10^{-5}	0.348
	2	3.0×10^{-4}	0.324	8.2×10^{-5}	0.187
	3	2.2×10^{-3}	0.141	3.1×10^{-4}	0.283

In a last step, I assess the model performance of the 3-layer model based on the absolute and relative errors of the determined error model fit (Table 5.7). Similar to the fitted error model of the 5-layer model, the error model fit of the 3-layer model approximates a linear error model with an intercept of zero. The relative error of the 3-layer model is with $b = 25\%$ slightly smaller than the relative error resulting from model calibration with the 5-layer model ($b = 27\%$). This indicates that the investigated aquifer may as well be represented by three instead of five aquifer segments.

Table 5.7: Calibration results of the locally anisotropic 3-layer model.

	3-layer model
$\mathbf{K_r^{eff}}[\mathbf{m/s}]$	1.7×10^{-3}
$\mathbf{K_z^{eff}}[\mathbf{m/s}]$	1.1×10^{-4}
$\vartheta[-]$	15.9
$\mathbf{K_{gp}}[\mathbf{m/s}]$	1.0×10^{0}
$\mathbf{K_{cl}}[\mathbf{m/s}]$	1.0×10^{-6}
$\mathbf{s_{ref,I}/s_{ref,II}/s_{ref,III}}$	0.039 / 0.036 / 0.030
$\mathbf{a}[\mathbf{m}]$	2.2×10^{-14}
$\mathbf{b}[-]$	0.25
$\mathbf{c}[\mathbf{m}]$	7.1×10^{-3}

5.7.6 Considering Water Extraction from a Single Well Screen

The results presented so far are based on joint inversion of pumping-test data in which water is extracted from different depths of the aquifer. Installing a partially penetrating well and sealing off the non-active well screens for extracting water from different depths, however, causes a considerable field effort. This raises the question of how much benefit is gained from tomographic pumping tests compared to tests in which water is extracted from a single well screen only when the target quantity is the hydraulic anisotropy. Towards this end, I inverted the data associated with hydraulic test I, II, and III separately and compared the results to the joint inversion of all three tests. Like before, I calibrated a locally anisotropic 1-layer model and a locally anisotropic model of multiple layers estimating the radial and vertical conductivities K_r and K_z of each horizontal layer, the conductivities of the gravel and clay fill K_{gp} and K_{cl}, and the drawdown at a virtual reference point s_{ref}. Based on the results presented in section 5.7.5, I consider a multi-layer model that consists of three horizontal layers (3-layer model).

Figure 5.16 shows a comparison of the measured and simulated drawdown s_{meas} and s_{sim} resulting from the 1-layer and 3-layer model calibration. The findings correspond to the results when considering the final error-model update in model calibration.
As discussed in section 5.4, the signal strength of drawdown varies among the three hydraulic tests, showing larger drawdown values when water was extracted from the middle screen than from the upper or lower screen. When inverting data from all three hydraulic tests jointly, the 1-layer model is not able to reproduce the observed differences in drawdown among the three hydraulic tests (for comparison see Figure 5.12c). In

contrast, the 1-layer model is able to meet the majority of measured drawdowns for all hydraulic tests when each test is considered independently in model calibration, but the estimated coefficients differ (Figure 5.16a-c).

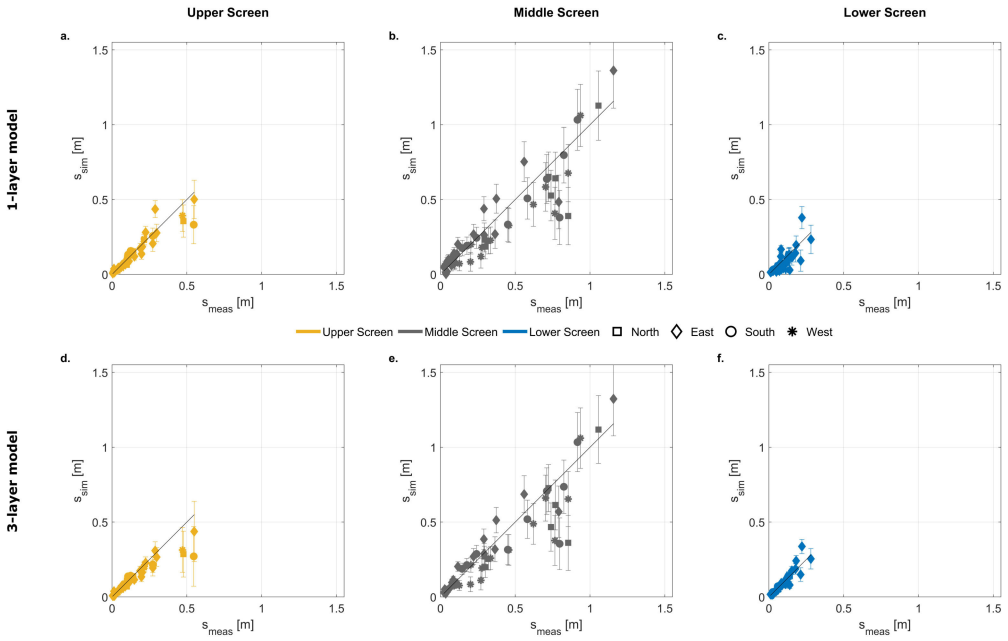

Figure 5.16: Field measurements versus simulated results of the 1-layer model when separately inverting data from hydraulic test I (a), II (b), and III (c) with errorbars according to the error model. Field measurements versus simulated results of the 3-layer model when separately inverting data from hydraulic test I (d), II (e), and III (f) with errorbars according to the error model. The black diagonal lines represent the 1:1 identity lines.

In case of the 3-layer model both the data inversion of pumping tests following the tomographic approach and the separate data inversion of each hydraulic test show a good fit between the simulated and measured drawdowns (Figure 5.12d and Figure 5.16d-f).

In a next step, I compare the estimated radial and vertical conductivities K_r and K_z of each horizontal layer available in the 1-layer and 3-layer model among the different inversion approaches.

K_r and K_z estimated with the 1-layer model correspond to the effective radial and vertical conductivities K_r^{eff} and K_z^{eff} and show similar values for the data inversion

of hydraulic test I and III (Table 5.8 and Figure 5.17a-b). In contrast, K_r^{eff} and K_z^{eff} resulting from data inversion of hydraulic test II are considerably smaller (Table 5.8 and Figure 5.17a-b). This confirms the assumption of vertical variations of hydraulic conductivity showing that the upper and lower part of the aquifer are more conductive than the middle part (for comparison see section 5.4 and section 5.7.2). For the inversion approach considering data from tomographic pumping tests the large drawdowns measured in hydraulic test II are highly erroneous so that they have little influence on the adjustment of hydraulic parameters. As a consequence, the estimated effective conductivities from data inversion considering the tomographic pumping tests resemble the estimates resulting from the separate data inversion of hydraulic test I and III (see Figure 5.17a-b).
Figure 5.17c and d display the standard deviations $\hat{\sigma}(lnK)$ of estimation associated with the estimated radial and vertical log-conductivities, respectively. All standard deviations of estimation are within an order of magnitude of 10^{-1} m and do not show any significant trend among the different inversion approaches.

Independent of the inversion approach, the 3-layer model estimates higher radial conductivities K_r for the upper and lower layer than for the middle layer (Figure 5.17e). In contrast, the vertical conductivities K_z show an increasing trend with increasing depth for all inversion approaches (Figure 5.17f). Despite the similar distribution of K_r and K_z among the different inversion approaches, the estimated parameters strongly differ. Consequently, also the resulting effective radial and vertical hydraulic conductivities K_r^{eff} and K_z^{eff} strongly differ (see Table 5.8) and range from $K_r^{eff} = 9.53 \times 10^{-4}$m/s (hydraulic test II) to $K_r^{eff} = 1.39 \times 10^{-2}$m/s (hydraulic test I), and $K_z^{eff} = 6.88 \times 10^{-7}$m/s (hydraulic test III) to $K_z^{eff} = 1.47 \times 10^{-4}$m/s (hydraulic test I). This leads to anisotropy ratios varying up to three orders of magnitude.
To examine the strong variations of estimated parameters, I consider the standard deviations $\hat{\sigma}(lnK)$ of estimation associated with the estimated radial and vertical log-conductivities (Figure 5.17g-h). In general, the standard deviations of estimation are significantly higher when inverting data from hydraulic test I, II, and III separately than when inverting the data from all three hydraulic tests jointly. The separate data inversions of hydraulic test I, II, and III show that the standard deviations of log-conductivities increase with increasing distance to the layer in which the active well screen is located. Thus, the estimated parameters are afflicted with higher uncertainties for layers that are located beyond the layer considered for water extraction.

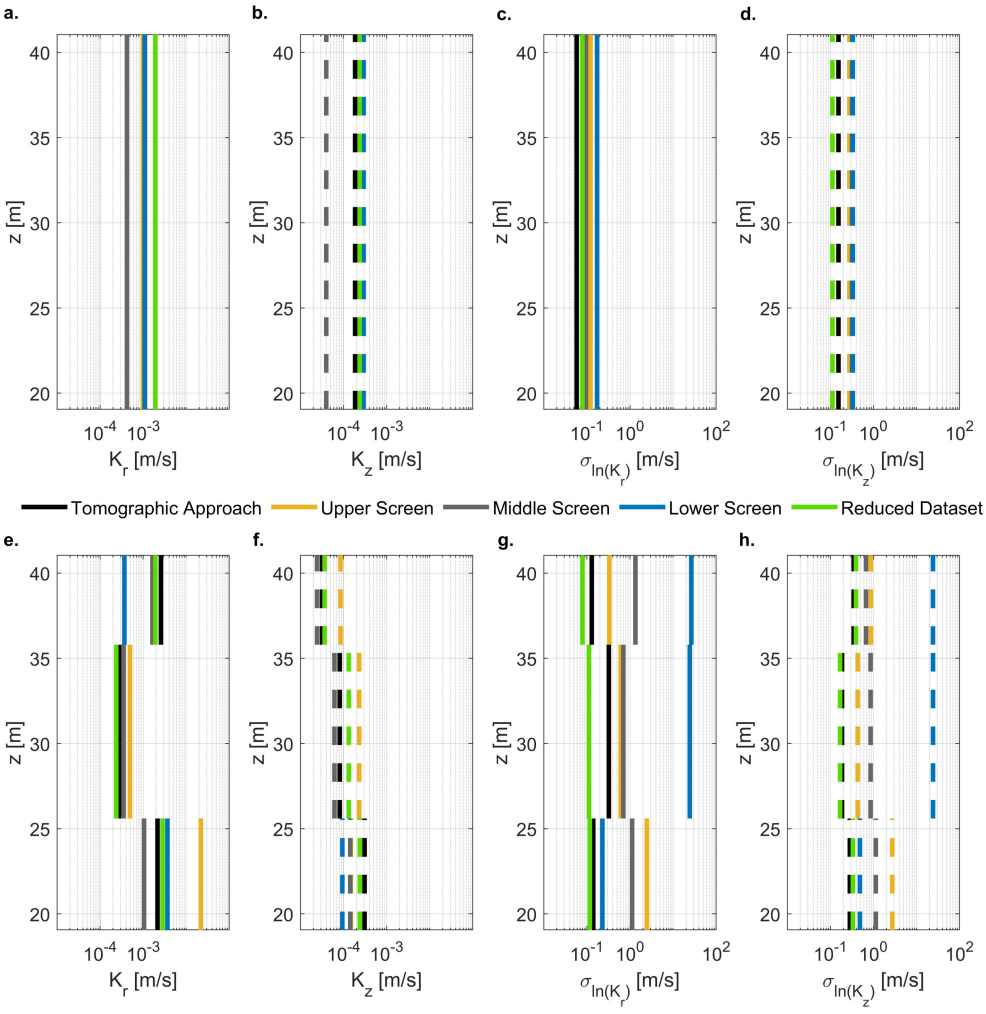

Figure 5.17: Radial and vertical conductivities K_r and K_z estimated for each available horizontal layer of the 1-layer model (a-b) and 3-layer model (e-f) considering the following inversion approaches: inverting data from pumping tests following the tomographic approach, inverting data from hydraulic test I (water extraction from the upper screen), II (water extraction from the middle screen), and III (water extraction from the lower screen), and inverting a reduced data set. Associated standard deviations of log-conductivities resulting from the 1-layer (c-d) and 3-layer model calibration (g-h) using different inversion approaches.

Finally, I assess the resulting error-model parameters (Table 5.8 and Table 5.9). The absolute errors of the 1-layer model show no significant abnormalities among the different inversion approaches. While the data inversion of pumping tests following the

Table 5.8: Calibration results of the locally anisotropic 1-layer model when separately inverting data of hydraulic test I, II, and III.

	Upper Screen	**Middle Screen**	**Lower Screen**
$\mathbf{K_r^{eff}}$[m/s]	9.9×10^{-4}	4.2×10^{-4}	1.1×10^{-3}
$\mathbf{K_z^{eff}}$[m/s]	2.4×10^{-4}	4.0×10^{-5}	2.9×10^{-4}
ϑ[−]	4.1	10.6	3.7
$\mathbf{K_{gp}}$[m/s]	4.0×10^{-1}	2.1×10^{-1}	2.4×10^{-1}
$\mathbf{K_{cl}}$[m/s]	1.2×10^{-8}	4.4×10^{-8}	1.4×10^{-8}
$\mathbf{s_{ref}}$	8.3×10^{-2}	2.3×10^{-1}	7.4×10^{-2}
a[m]	6.5×10^{-3}	2.3×10^{-2}	2.2×10^{-14}
b[−]	0.32	0.20	0.35
c[m]	2.5×10^{-1}	6.9×10^{-6}	1.1×10^{-2}

tomographic approach has shown a considerably high relative error of $b = 88\%$ (see Table 5.4), the separate data inversions of hydraulic test I, II, and III reveal smaller relative errors ranging between 20% and 35% (Table 5.8).
The absolute errors for the separate data inversions of hydraulic test I, II, and III with the 3-layer model are within the mm-range (Table 5.9). While the relative errors are similar for the separate data inversion of hydraulic test II and the inversion of data considering the tomographic approach, the relative error of the data inversion of hydraulic test III is twice as high. Inverting data from hydraulic test I reveals a very large relative error of $b = 296$ and a physically unreasonable coefficient c of 453 m.

Table 5.9: Calibration results of the locally anisotropic 3-layer model when separately inverting data of hydraulic test I, II, and III.

	Upper Screen	**Middle Screen**	**Lower Screen**
$\mathbf{K_r^{eff}}$[m/s]	1.4×10^{-2}	9.5×10^{-4}	2.3×10^{-3}
$\mathbf{K_z^{eff}}$[m/s]	1.5×10^{-4}	7.3×10^{-5}	6.9×10^{-7}
ϑ[−]	94.5	13.0	3382.2
$\mathbf{K_{gp}}$[m/s]	8.0×10^{-2}	5.1×10^{-1}	2.3×10^{-4}
$\mathbf{K_{cl}}$[m/s]	1.8×10^{-9}	1.0×10^{-6}	3.7×10^{-9}
$\mathbf{s_{ref}}$	1.0×10^{-8}	8.1×10^{-2}	1.0×10^{-8}
a[m]	5.0×10^{-3}	7.5×10^{-3}	1.4×10^{-3}
b[−]	296	0.21	0.43
c[m]	454	7.2×10^{-6}	2.3×10^{-1}

Overall, the 1-layer model shows a better agreement between the simulated and measured drawdowns for the data inversion from pumping tests with a single well screen.

Also, the resulting relative errors are smaller compared to the data inversion considering the tomographic pumping-test approach. The conductivity estimates resulting from the separate data inversions, however, differ for the different hydraulic tests. Thus, considering pumping tests with a single well screen can lead to the wrong interpretation that the aquifer is anisotropic but vertically homogeneous.

Also the 3-layer model reveals different parameter estimates for pumping tests with a single well screen which strongly depend on the chosen well screen. The associated standard deviations of estimation indicate that including a single well screen yields acceptable uncertainties only for the estimated conductivities of the layer in which the active well screen is located. Therefore, it only makes sense to calibrate a multi-layer model when water is extracted from different depths and the data is jointly considered in the inversion.

5.7.7 Reducing the Number of Observation Points

The field effort of the tomographic pumping tests is not only related to the number of extraction screens of the pumping well but also to the number of observation wells. In the following, I therefore assess the data inversion of tomographic pumping tests considering a reduced dataset. The reduced dataset is based on drawdown measurements from observation wells in one direction to the pumping well only. I calibrated the locally anisotropic 1-layer and 3-layer model, considering all field observations from wells that are located east to the pumping well since these are the observation wells that reveal the largest number of depth-oriented observation points.

Figure 5.18a and b show that the 1-layer model and 3-layer model reproduce the true measurements with a reduced dataset similarly well as with the complete dataset (for comparison see Figure 5.12c and Figure 5.15b). For the reduced dataset the 3-layer model even reveals a distinctively good fit. This can be attributed to the small influence of potential local heterogeneities and horizontal anisotropy when field measurements are considered from one direction only.

The estimated radial and vertical conductivities K_r and K_z from the 1-layer and 3-layer model calibrations with a reduced dataset are similar to the results of the 1-layer and 3-layer model calibrations with the complete dataset (compare Table 5.10 with Table 5.3 and Table 5.6). Consequently, also the effective radial and vertical conductivities,

and the resulting anisotropy ratios are comparable. Additionally, the standard deviations of the log-conductivities are very similar for the reduced and complete dataset, for both the 1-layer (Figure 5.17c-d) and 3-layer model (Figure 5.17g-h).

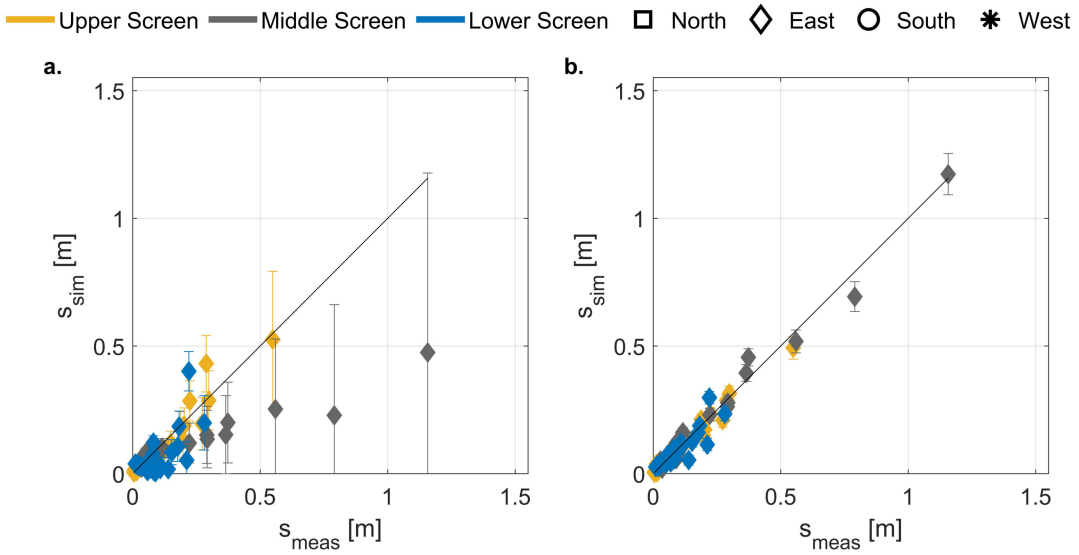

Figure 5.18: Field measurements versus simulated results of the 1-layer model (a) and 3-layer model (b) considering data inversion of tomographic pumping tests with a reduced dataset and with errorbars according to the error model. The black diagonal lines represent the 1:1 identity lines.

Finally, I consider the parameters of the error model resulting from the 1-layer and 3-layer model calibration with a reduced data set (Table 5.10). For both the 1-layer and 3-layer model the absolute errors are higher for the reduced dataset ($a = 1.6 \times 10^{-2}$ m and $a = 1.1 \times 10^{-2}$ m) than for the complete dataset ($a = 3.1 \times 10^{-3}$ m and $a = 2.2 \times 10^{-14}$ m). In contrast, the relative errors are reduced from $b = 0.88$ for the complete dataset to $b = 0.81$ for the reduced dataset with the 1-layer model, and from $b = 0.25$ (complete dataset) to $b = 0.06$ (reduced dataset) with the 3-layer model.

Overall, the results considering a reduced dataset do not show significant differences to the results based on the complete dataset. This suggests that installing observation wells in one direction to the pumping well only would have been sufficient. Yet, to identify horizontal heterogeneity and anisotropy in the first place, individual observation points in different directions to the pumping well are required.

Table 5.10: Calibration results of the locally anisotropic 1-layer and 3-layer model considering a reduced dataset.

	1-layer model	**3-layer model**
$\mathbf{K_r^{eff}}[\mathbf{m/s}]$	1.0×10^{-3}	2.1×10^{-3}
$\mathbf{K_z^{eff}}[\mathbf{m/s}]$	2.1×10^{-4}	1.2×10^{-4}
$\vartheta[-]$	4.8	16.5
$\mathbf{K_{gp}}[\mathbf{m/s}]$	1.9×10^{-3}	7.2×10^{-3}
$\mathbf{K_{cl}}[\mathbf{m/s}]$	2.3×10^{-7}	1.0×10^{-6}
$\mathbf{s_{ref,I}/s_{ref,II}/s_{ref,III}}$	0.080 / 0.069 / 0.131	0.038 / 0.032 / 0.018
$\mathbf{a}[\mathbf{m}]$	1.6×10^{-2}	1.1×10^{-2}
$\mathbf{b}[-]$	0.81	0.06
$\mathbf{c}[\mathbf{m}]$	4.2×10^{-1}	4.8×10^{-12}

Chapter 6

Conclusions and Outlook

This thesis was motivated by the shortage of hydrogeological methods on estimating the hydraulic anisotropy interrelated with the vertical variability of hydraulic conductivity. Knowledge of differences in the horizontal and vertical hydraulic conductivities is crucial for many hydrogeological applications, especially where a strong vertical flow component is present, e.g. in the design of horizontal collector wells, in riverbank-filtration setups or in the design of geotechnical dewatering systems. The main goal of this thesis was to investigate the potential of a field method similar to that of hydraulic tomography for estimating the large-scale hydraulic anisotropy in stratified aquifers induced by the vertical heterogeneity on smaller scales. The approach is based on calibrating groundwater flow models using data of sequential pumping tests with partially penetrating wells, during which water is extracted from different aquifer sections and the hydraulic response is measured at different radial and vertical distances to an extraction screen.

In the following, I summarize the main findings of this work and draw conclusions regarding the research questions and goals defined in section 1.1.

Should data be collected based on a predefined model or vice versa?
To address the challenge of model selection and data collection, I developed an optimization framework in which both challenges are jointly addressed. The proposed framework is based on mimicking best-fit model calibration using simulated and perturbed measurements in an ensemble of virtual realities and selecting the simulated measurements of ensembles of reasonable model candidates. In addition, the calibration is performed for a variety of feasible data-acquisition strategies defined in different

observation designs. The approach was tested for the application case of radial flow in stratified aquifers, for which the virtual realities exhibited a high resolution of the vertical hydraulic-conductivity profiles, while the model candidates proposed considered 1, 2, 3, or 6 horizontal layers of uniform, anisotropic conductivity. The conceptual question of the model selection was therefore whether one should estimate macroscopic hydraulic anisotropy, or hydraulic heterogeneity, or a combination of both, whereas the measurement-design question was whether multi-level piezometers should be placed particularly close to the pumping well (where drawdown signals are strong and exhibit strong vertical differences but are particularly prone to spatial misplacements of the piezometers) or further away (where both the signal and systematic measurement errors are smaller).

Instead of judging the calibrated conceptual models and measurement designs by the ability of the models to reproduce the measurements in calibration, I defined the following prediction targets of the calibrated models: the effective radial and vertical hydraulic conductivities and the dewatering flux of a hypothesized construction pit. The prediction targets of the models should be as close to the predictions of the virtual realities as possible.

The most complex conceptual model performed best with respect to meeting the "true" measurements during calibration. Choosing the number of layers to be identical to the number of observation depths was advantageous, since this selection not only led to the best fit of observations but also to the best prediction of the effective vertical hydraulic conductivity and the dewatering flux necessary to keep the hypothesized construction pit dry. For the effective horizontal conductivity chosen as target quantity, the superiority of the 6-layer model was less persistent over the different observation designs.

Quite clearly, ignoring the observation design in model selection would deteriorate the model performance. For example, placing observation wells at far radial distances to the pumping well hampers resolving hydraulic-conductivity values of several layers. Likewise, ignoring the model choice when optimizing the measurement strategy would make little sense. The analysis also showed that the joint selection of the conceptual model and the measurement strategy depends on the ultimate objective of the model application. Reducing the prediction uncertainty in the effective horizontal conductivity required observation points with large radial distance, whereas reducing the prediction

uncertainty in the effective vertical conductivity and the dewatering flux required multi-level observation wells at small radial distances, where the vertical hydraulic gradients are the biggest.

Design a well network to resolve hydraulic anisotropy

While the study of jointly optimizing measurement and modeling strategies was based on hypothetical field scenarios and synthetic data only, I also investigated a field application of the hydraulic tests. To perform pumping tests with a tomographic setup a new research site was established in a fluvial gravel aquifer in the Upper Rhine Valley. At the site, a well network was installed with a large diameter pumping well in the center of the network and observation wells were placed at different radial distances to the pumping well. The pumping well was designed with three isolated well screens to enable water extraction from different depths. By extracting water from a partially penetrating well, a strong vertical flow component is induced which is required to resolve the directional dependence of hydraulic conductivity in stratified aquifers. The observation wells were installed either as multi-level wells or as clusters of partially penetrating wells, facilitating the observation of hydraulic responses at different depths. To account for horizontal anisotropy, at least a few observation wells need to be placed in a different direction to the pumping well than along a main transect only.

How should large field datasets be handled and which data compression should be used before inversion?

A total of 22 pumping-test series were performed in which water was sequentially extracted from the three different well screens, measuring more than 1000 transient drawdown responses in the well network of 58 observation points. By repeating the tests with different pumping rates and then rescaling the results to a common rate allowed checking the reproducibility of the performed pumping tests. Considering extraction from each well screen at a time, I averaged all reproducible data of an observation point reducing the large data volume to a manageable size.

Analyzing the steady-shape regime of the pumping tests provided a beneficial approach in data analysis, with which the challenges related to analyzing transient data or of reaching steady state drawdown in field applications were avoided (Bohling et al. 2002, Bohling et al. 2007). In posterior data calibration, I have demonstrated that implementing a virtual reference point to compute drawdown differences is a rea-

sonable alternative to the computation of drawdown differences based on pairs of true observation points, for which inherent measurement errors might be propagated.

How should a groundwater flow model be setup to represent well the true system and how should model performance be evaluated?
To assess the effects of different model settings, I fitted the response of an anisotropic single-layered homogeneous model, and a locally anisotropic multi-layer model to the field data of the pumping tests with water extraction from different depths. To setup the multi-layer model, I used available lithologic information from the drilling profile of the pumping well. In addition to calibrate radial and vertical hydraulic conductivity for each horizontal layer, I calibrated the error model parameters for assessing model performance.

The results showed that if sufficient data are available, it is preferable to resolve the main vertical structure of hydraulic conductivity over fitting a uniform effective conductivity tensor. A better agreement was obtained between simulated and measured drawdown, and systematic bias was avoided, with the multi-layer model than with the single-layer one. The resulting effective radial and vertical conductivities of the multi-layer model turned out to be informative only when the vertical variability in hydraulic conductivity within the aquifer was considered.

In general, the determination of locally anisotropic horizontal layers could be validated by comparing the locally anisotropic groundwater models to poorly performing locally isotropic groundwater models. The locally anisotropic multi-layer model had a better overall performance and was further enhanced by adapting the number of horizontal layers to those indicated by the calibration results, showing that a multi-layer model with one layer per well screen is advantageous. To define the position of the layer boundaries, the initially implemented information from the drilling profile proved to be valuable, but assuming more layers than extraction depths was not really necessary at the site.

Finally, I compared the data inversion of pumping tests with water extraction from different depths to an approach in which data associated with water extraction from a single well screen was inverted. When analyzing data using a partially pumping well with a single well screen, anisotropic one-layer models could be fitted to the data reasonably well, but the coefficients highly depended on the depth of water extraction. The ability of reproducing the data could easily be misinterpreted as having understood the

hydraulic structure of the aquifer. The results showed that a combination of pumping tests with water extraction from different depths is necessary to estimate the vertical variability and anisotropy of stratified aquifers.

6.1 Recommendations

I have demonstrated the applicability of inverting tomographic pumping tests that target the vertical variability and anisotropy of potentially stratified aquifers. Of course, the experimental effort of installing a multi-section partially penetrating well of a large diameter comes with a considerably high effort. Performing the pumping tests, however, is not restricted to such a large-diameter well or to the screen lengths considered in this study. The well was designed to allow a pumping rate that induces a sufficiently large depression cone within a measurable signal range in a highly productive gravel aquifer. For other applications, the diameter and screen lengths of the extraction well could be adjusted depending on the scale and aquifer under investigation.

To justify the assumption of radial symmetric flow (neglecting horizontal heterogeneity and/or anisotropy) it was important to install observation wells in several directions from the pumping well. To further improve the experimental design, I recommend to tailor the installation of multi-level observation wells. The efforts required to install multi-level observation wells, to observe the hydraulic response in those wells, and to process the obtained field data may be reduced by developing a depth-oriented measurement network particularly in one direction while considering individual control points in further directions.

As mentioned previously, the multi-layer model considered for the inversion of the field test data was based only on geological information obtained through the drilling profile of the pumping well. To improve the delineation of hydraulically relevant layers, I suggest performing flowmeter and direct-push injection-logging tests. With this, the existence of consistent layers with higher or lower conductivities across several vertical profiles could be verified.

Overall, the hydraulic-tomography approach investigated in this thesis shows its potential for estimating the hydraulic anisotropy interrelated with the vertical heterogeneity in stratified aquifers. Yet, the approach requires precise knowledge on vertical

variations in hydraulic conductivity. Further research is necessary to reduce the uncertainty of anisotropy estimates by improving the integration of subsurface heterogeneity. I recommend to combine the approach with further hydrogeological subsurface investigation techniques. For example, using tracer tomography to delineate preferential flow paths can provide valuable information regarding vertical conductivity variations. With this, model errors might be reduced and the prediction uncertainty of anisotropy could be decreased. The same monitoring network used for the hydraulic tomography could be used to monitor the evolution of the tracer plume, minimizing the extra field work.

Even though the horizontal variability is often smaller than the differences among the vertical layers, the assumption of two-dimensional radial symmetric flow may not always be valid at other field sites and a three dimensional problem might have to be considered.

References

Aigner, T., Heinz, J., Hornung, J., and Asprion, U. (1998). A hierarchical process-approach to reservoir heterogeneity: Examples from outcrop analogues. Bull. Cent. Rech. Elf Explor. Prod. 22, 1-11.

Bair, E. S. and Lahm, T. D. (1996). Variations in capture-zone geometry of a partially penetrating pumping well in an unconfined aquifer. *Ground Water*, 34(5):842–852.

Barlebo, H. C., Hill, M. C., and Rosbjerg, D. (2004). Investigating the macrodispersion experiment (made) site in columbus, mississippi, using a three-dimensional inverse flow and transport model. *Water Resources Research*, 40(4).

Bear, J. (1972). *Dynamics of Fluids in Porous Media*. Dover Publications, Inc., New York.

Bennett, J. P., Haslauer, C. P., and Cirpka, O. A. (2017). The impact of sedimentary anisotropy on solute mixing in stacked scour-pool structures. *Water Resources Research*, 53(4):2813–2832.

Bennett, J. P., Haslauer, C. P., Ross, M., and Cirpka, O. A. (2019). An open, object-based framework for generating anisotropy in sedimentary subsurface models. *Groundwater*, 57(3):420–429.

Beven, K. (2002). Towards a coherent philosophy for modelling the environment. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 458(2026):2465–2484.

Bohling, G. C. (2009). Sensitivity and resolution of tomographic pumping tests in an alluvial aquifer. *Water Resources Research*, 45:Artn W02420.

Bohling, G. C., Butler, J. J., Zhan, X. Y., and Knoll, M. D. (2007). A field assessment of the value of steady shape hydraulic tomography for characterization of aquifer heterogeneities. *Water Resources Research*, 43(5).

Bohling, G. C., Zhan, X. Y., Butler, J. J., and Zheng, L. (2002). Steady shape analysis of tomographic pumping tests for characterization of aquifer heterogeneities. *Water Resources Research*, 38(12).

Borghi, A., Renard, P., and Courrioux, G. (2015). Generation of 3d spatially variable anisotropy for groundwater flow simulations. *Groundwater*, 53(6):955–958.

Bourdet, D., Ayoub, J., and Pirard, Y. (1989). Use of Pressure Derivative in Well Test Interpretation. *SPE Formation Evaluation*, 4(02):293–302.

Butler, J. J., Dietrich, P., Wittig, V., and Christy, T. (2007). Characterizing hydraulic conductivity with the direct-push permeameter. *Ground Water*, 45(4):409–419.

Butler, J. J., Healey, J. M., McCall, G. W., Garnett, E. J., and Loheide, S. P. (2002). Hydraulic tests with direct-push equipment. *Ground Water*, 40(1):25–36.

Cardiff, M. and Barrash, W. (2011). 3-d transient hydraulic tomography in unconfined aquifers with fast drainage response. *Water Resources Research*, 47:Artn W12518.

Cardiff, M., Barrash, W., Kitanidis, P. K., Malama, B., Revil, A., Straface, S., and Rizzo, E. (2009). A potential-based inversion of unconfined steady-state hydraulic tomography. *Ground Water*, 47(2):259–270.

Chaloner, K. and Verdinelli, I. (1995). Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273–304.

Chen, X. H., Burbach, M., and Cheng, C. (2008). Electrical and hydraulic vertical variability in channel sediments and its effects on streamflow depletion due to groundwater extraction. *Journal of Hydrology*, 352(3-4):250–266.

Chen, X. H., Song, J. X., and Wang, W. K. (2010). Spatial variability of specific yield and vertical hydraulic conductivity in a highly permeable alluvial aquifer. *Journal of Hydrology*, 388(3-4):379–388.

Chiogna, G., Cirpka, O. A., Rolle, M., and Bellin, A. (2015). Helical flow in three-dimensional nonstationary anisotropic heterogeneous porous media. *Water Resources Research*, 51(1):261–280.

Cirpka, O. A., Chiogna, G., Rolle, M., and Bellin, A. (2015). Transverse mixing in three-dimensional nonstationary anisotropic heterogeneous porous media. *Water Resources Research*, 51(1):241–260.

Coleman, T. F. and Li, Y. Y. (1996). An interior trust region approach for nonlinear minimization subject to bounds. *Siam Journal on Optimization*, 6(2):418–445.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39(1):1–38.

Dietrich, C. R. and Newsam, G. N. (1993). A fast and exact method for multidimensional gaussian stochastic simulations. *Water Resources Research*, 29(8):2861–2869.

Dietrich, P., Butler, J. J., and Faiss, K. (2008). A rapid method for hydraulic profiling in unconsolidated formations. *Ground Water*, 46(2):323–328.

Doherty, J. and Christensen, S. (2011). Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resources Research*, 47.

Einarson, M. D. and Cherry, J. A. (2002). A new multilevel ground water monitoring system using multichannel tubing. *Ground Water Monitoring and Remediation*, 22(4):52–65.

Erdal, D. and Cirpka, O. A. (2019). Global sensitivity analysis and adaptive stochastic sampling of a subsurface-flow model using active subspaces. *Hydrology and Earth System Sciences*, 23(9):3787–3805.

Freeze, R. and Cherry, J. (1979). *Groundwater*. 0-13-365312-9. Prentice-Hall.

Gianni, G., Doherty, J., and Brunner, P. (2019). Conceptualization and calibration of anisotropic alluvial systems: Pitfalls and biases. *Groundwater*, 57(3):409–419.

Gottlieb, J. and Dietrich, P. (1995). Identification of the permeability distribution in soil by hydraulic tomography. *Inverse Problems*, 11(2):353–360.

Hill, M. C. and Tiedeman, C. R. (2007). *Effective groundwater model calibration with analysis of data, sensitivities, and uncertainty*. Hoboken, New Jersey: John Wiley and Sons.

Hoge, M., Wohling, T., and Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, 54(3):1688–1715.

Klammler, H., Hatfield, K., Nemer, B., and Mathias, S. A. (2011). A trigonometric interpolation approach to mixed-type boundary problems associated with permeameter shape factors. *Water Resources Research*, 47:Artn W03510.

Klammler, H., Layton, L., Nemer, B., Hatfield, K., and Mohseni, A. (2017). Theoretical aspects for estimating anisotropic saturated hydraulic conductivity from in-well or direct-push probe injection tests in uniform media. *Advances in Water Resources*, 104:242–254.

Kruseman, G. and de Ridder, N. (1994). *Analysis and Evaluation of Pumping Test Data*, volume 47. International Institute for Land Reclamation and Improvement, Wageningen, The Netherlands.

Lessoff, S. C., Schneidewind, U., Leven, C., Blum, P., Dietrich, P., and Dagan, G. (2010). Spatial characterization of the hydraulic conductivity using direct-push injection logging. *Water Resources Research*, 46:Artn W12502.

Leube, P. C., Geiges, A., and Nowak, W. (2012). Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. *Water Resources Research*, 48.

LGRB (2004). Geological map of Baden-Württemberg, Germany. *Landesamt für Geologie, Rohstoffe und Bergbau Baden-Württemberg*.

Maier, R., Gonzalez-Nicolas, A., Leven, C., Nowak, W., and Cirpka, O. A. (2020). Joint optimization of measurement and modeling strategies with application to radial flow in stratified aquifers. *Water Resources Research*, 56(7):ARTN e2019WR026872.

Maier, R., Strasser, D., Cirpka, O. A., and Leven, C. (2021). Data set on pumping tests in a partially penetrating well to resolve vertical aquifer heterogeneity and hydraulic anisotropy. Eberhard Karls Universität Tübingen. http://hdl.handle.net/10900.1/274eabc9-0edd-40b1-9e34-541c09c4479c.

Müller, P. (2005). Simulation Based Optimal Design. In Dey, D. K. and Rao, C. R., editors, *Handbook of Statistics*, volume 25 of *Bayesian Thinking*, pages 509–518. Elsevier.

Oreskes, N. (1998). Evaluation (not validation) of quantitative models. *Environmental Health Perspectives*, 106:1453–1460.

Paradis, D., Gloaguen, E., Lefebvre, R., and Giroux, B. (2015). Resolution analysis of tomographic slug test head data: Two-dimensional radial case. *Water Resources Research*, 51(4):2356–2376.

Paradis, D., Gloaguen, E., Lefebvre, R., and Giroux, B. (2016). A field proof-of-concept of tomographic slug tests in an anisotropic littoral aquifer. *Journal of Hydrology*, 536:61–73.

Paradis, D. and Lefebvre, R. (2013). Single-well interference slug tests to assess the vertical hydraulic conductivity of unconsolidated aquifers. *Journal of Hydrology*, 478:102–118.

Poeter, E. P. and Hill, M. C. (1997). Inverse models: A necessary next step in groundwater modeling. *Ground Water*, 35(2):250–260.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163.

RamaRao, B. S., LaVenue, A. M., De Marsily, G., 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.

Renard, P., Glenz, D., and Mejias, M. (2009). Understanding diagnostic plots for well-test interpretation. *Hydrogeology Journal*, 17(3):589–600.

Rubin, Y. (2003). *Applied Stochastic Hydrogeology*. Oxford University Press, New York.

Sanchez-Leon, E., Leven, C., Haslauer, C. P., and Cirpka, O. A. (2016). Combining 3d hydraulic tomography with tracer tests for improved transport characterization. *Groundwater*, 54(4):498–507.

Stoll, M. (2020). Assessing groundwater in the Rhine Valley - Numerical inversion of a step-drawdown pumping test for estimating hydraulic anisotropy. Master thesis, Eberhard Karls Universität Tübingen.

Tiedeman, C. R., Hill, M. C., D'Agnese, F. A., and Faunt, C. C. (2003). Methods for using groundwater model predictions to guide hydrogeologic data collection, with application to the death valley regional groundwater flow system. *Water Resources Research*, 39(1).

Vanlier, J., Tiemann, C. A., Hilbers, P. A. J., and van Riel, N. A. W. (2014). Optimal experiment design for model selection in biochemical networks. *Bmc Systems Biology*, 8.

Walker, W. E., Harremoës, P., Rotmans, J., Sluijs, J. P. v. d., Asselt, M. B. A. v., Janssen, P., and Krauss, M. P. K. v. (2003). Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment*, 4(1):5–17.

Wirsing, G. and Luz, A. (2007). Hydrogeologischer Bau und Aquifereigenschaften der Lockergesteine im Oberrheingraben (Baden-Württemberg). *Landesamt für Geologie, Rohstoffe und Bergbau Baden-Württemberg*.

Woessner, W. W. and Anderson, M. P. (1996). Good model bad model, understanding the flow modeling process. *Subsurface Fluid-Flow (Ground-Water and Vadose Zone) Modeling*, 1288:14–23.

Yeh, T. C. J. and Liu, S. Y. (2000). Hydraulic tomography: Development of a new aquifer test method. *Water Resources Research*, 36(8):2095–2105.

Yeh, W. W.-G. and Yoon, Y. S. (1981). Aquifer parameter identification with optimum dimension in parameterization. *Water Resources Research*, 17(3):664–672.

Zlotnik, V. and Ledder, G. (1996). Theory of dipole flow in uniform anisotropic aquifers. *Water Resources Research*, 32(4):1119–1128.

Zschornack, L., Bohling, G. C., Butler, J. J., and Dietrich, P. (2013). Hydraulic profiling with the direct-push permeameter: Assessment of probe configuration and analysis methodology. *Journal of Hydrology*, 496:195–204.