Hydraulic Tomography in Unconfined Aquifers and the Importance of Geological Data: Laboratory and Field Studies

by

Zhanfeng Zhao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Earth Sciences

Waterloo, Ontario, Canada, 2017

© Zhanfeng Zhao 2017

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Abstract

This dissertation first demonstrates the effectiveness of hydraulic tomography (HT) in characterizing the heterogeneity of both the saturated and unsaturated zone parameters through a laboratory unconfined aquifer. Specifically, a geostatistical inversion algorithm that considers variably saturated flow processes (successive linear estimator (SLE) developed by *Mao et al.* (2013c))) in unconfined aquifers is utilized to obtain tomograms of hydraulic conductivity (K), specific storage (S_s), and the unsaturated zone parameters (pore size parameter (α) and saturated water content (θ_s)) for the Gardner-Russo's model. The estimated tomograms accurately capture the locations of heterogeneity including high and low K layers within the saturated and unsaturated zones compared to the true stratigraphy visible in the sandbox, as well as reasonable distribution patterns of α and θ_s for the Gardner-Russo's model. The estimated tomograms are then used to predict drawdown responses of pumping tests not used in the inverse modeling effort. The simulated and observed drawdown curves show an excellent agreement for observations in both the saturated and unsaturated zones. Drawdown predictions of the geostatistical inversion approach are significantly better than those based on the homogeneous assumption. Results of this study demonstrates the robust performance of HT that considers variably saturated flow processes in unconfined aquifers and the unsaturated zone above it, and substantiate the unbiased and minimal variance of HT analysis with the SLE algorithm.

Then, this dissertation investigates the importance of geological data in HT through

sandbox experiments. Four geological models with homogeneous units are constructed with borehole data of varying accuracy. These geological models are calibrated to multiple pumping test data of two different pumping and observation densities. Results show that both accurate and inaccurate geological models can be well calibrated and inaccurate geological models yield poor drawdown predictions for model validation. Moreover, model calibration and validation comparisons among layer-based geological models and a highly parameterized geostatistical model show that the performance gap between the approaches decreases as the number of pumping tests and monitoring locations are reduced. Next, four geological models are populated with permeameter test K values for each layer and used as prior mean information in geostatistical inverse models. Results show that the estimated K tomograms preserve geological features especially in areas where drawdown data are not available. Overall, this sandbox study emphasizes that accurate geological data is important for incorporating into HT surveys when data from pumping tests are sparse.

Finally, this dissertation looks into the importance of geological data in HT through field experiments conducted at a highly heterogeneous glaciofluvial deposit at the North Campus Research Site (NCRS) of the University of Waterloo. Unlike the sandbox study in which the stratigraphy is perfectly known, geological data are obtained from 18 boreholes at NCRS. Two geological models of different resolutions are constructed. One model contains 19 layers while the other model merges some of the units resulting in five layers. Steady state pressure head data of 14 pumping tests are selected from the site for model calibration and validation purposes. The results are first compared with permeameter-estimated K profiles along boreholes. Results reveal that the simultaneous calibration of geological models to seven pumping test data yields K values that correctly reflect the general patterns of vertical distributions of permeameter-estimated K. In addition, this study finds that using

a geological model as prior information in the geostatistical inversion approach leads to improved correspondence of K estimates to permeameter test results along wells, as well as in preserving geological features where drawdown measurements are lacking. Therefore, the field study suggests incorporating geological data for HT analysis based on geostatistical inverse modeling approaches when reliable geological data are available.

Acknowledgements

This dissertation was completed under different kinds of support and encouragements from many people. First of all, I would like to thank my supervisor Dr. Walter A. Illman for his constructive guidance and great support for the past four years and for providing me the opportunity of pursuing this Ph.D. at the University of Waterloo. Dr. Walter A. Illman has been very patient and thoughtful for research problems I have encountered. The experience and knowledge I have learnt from him would be greatly beneficial for future works.

Many thanks to my committee members: Dr. Dave L. Rudolph, Dr. André Unger, and Dr. Steve J. Berg for their help and discussions which contributed to this work. Special thanks to Dr. Steve J. Berg for many useful suggestions on both technical and field works, which allowed me to complete the research more efficiently and successfully. Thanks also go out to my internal-external examiners Dr. Bryan Tolson and Dr. Brent Sleep for their valuable time to review this work. Special thanks to research technician Wayne Noble for his support on field work.

Finally, I thank my parents and family members in China for their endless supports.

Table of Contents

List of Tables

List of Figures

Introduction

1.1 Review of Hydraulic Tomography

Groundwater investigations and exploitation rely on the accurate characterization of subsurface conditions. Pumping tests are extensively carried out at areas of interest to probe the responses and efforts are put in to obtain valid hydraulic parameters for future predictions of various purposes, like coal mine exploitation and development of municipal water supply. Based on the simplification of homogeneous aquifer assumption, type curves (e.g., *Theis*, 1935; *Neuman*, 1972, 1974) can be utilized to estimate the hydraulic conductivity and specific storage of the aquifer. However, hydrogeological issues like contaminant transport also rely on the accurate characterization of aquifer heterogeneity especially at finer scales. In order to deal with this issue, various approaches to build geological models based on the knowledge of either large scale spatial or local scale distribution have been developed (e.g., *Koltermann and Gorelick*, 1996; *de Marsily et al.*, 2005).

At the local scale, when point measurements are available, traditional geostatistical method such as kriging can obtain a high resolution hydraulic conductivity map based on sample locations. The main criticism of kriging is that the interpolated maps may be too smooth and not represent heterogeneity accurately. To overcome this limitation, indicator kriging (*Journel and Isaaks*, 1984; *Journel and Gomez-Hernandez*, 1993) has been developed to include the irregular geometry of geological bodies, allowing for representation of

main geometric patterns in space. Similarly, the Markov Chain approach (e.g. *Carle and Fogg*, 1997; *Weissmann et al.*, 1999), based on transition probability, directly interpolates material categories to generate maps including soft information such as geologic stratigraphy. Alternatively, multiple point geostatistics (MPG) (*Strebelle*, 2002; *Blouin et al.*, 2013) that uses multi-point statistics extracted directly from training images can generate subsurface heterogeneity maps (*Comunian et al.*, 2011).

More complex methods like genesis models (e.g., *Koltermann and Gorelick*, 1996; *Teles et al.*, 2004) mathematically simulate geological processes that create the medium although this approach requires a very large modeling domain to generate the deposits. Such a large domain may not be practical for typical hydrogeological models.

Generally, the developed methods evolve to produce geological models that are more realistic so that geological variability can be simulated at different scales. Then, these models are either assigned with assumed hydraulic parameter values or calibrated to available hydraulic head data.

At the regional scale, a commonly adopted approach is the construction of deterministic geological models based on available geological information (e.g., *Refsgaard et al.*, 2012; *Troldborg et al.*, 2007). These models are then calibrated to hydraulic head data under steady-state or transient conditions. Due to our limited knowledge of geology and conceptualization uncertainty, a single geological model tends to produce biased predictions (*Troldborg et al.*, 2007). In order to consider the impact of model uncertainty, alternative models are suggested by *Refsgaard et al.* (2012). When calibrated to only one kind of data set, different conceptual models may perform equally well during the calibration process but could yield quite significant differences in terms of prediction (*Rojas et al.*, 2008; *Troldborg et al.*, 2007). Thus, *Rojas et al.* (2010) suggests that more data (e.g., pressure

heads, permeability measurements) will likely better discriminate different geological models. *Harrar et al.* (2003) concluded that the structure of a geological model poses different optimized values for individual zones, but when calibrating using sufficient data one can constrain the inverse models so that the mean distribution of hydraulic conductivity of the entire domain is well represented. Moreover, errors originating from model construction will be compensated by biased parameter estimates during the optimization process (i.e., model calibration, see *Troldborg et al.*, 2007). Evaluation of model uncertainty by *Rojas et al.* (2008, 2010) revealed that the conceptual uncertainty will be a dominant source of prediction uncertainty once models are used for predictions beyond the data used for calibration. So far, experience from various previous studies show that deterministic geological model uncertainty at regional scales can neither be avoided nor eliminated.

On the other hand, parameter estimation using either homogeneous or heterogeneous conceptual models with a limited spatial observation of drawdowns induced by a single pumping test may yield scenario dependent effective parameters as pointed out by *Wu et al.* (2005); *Straface et al.* (2007); *Wen et al.* (2010); *Huang et al.* (2011); *Berg and Illman* (2011b, 2013, 2015), and *Sun et al.* (2013). This scenario dependent behavior can be attributed to the dependence of the estimated parameter on the heterogeneity around the pumping location. Thus, new approaches have to be developed to address the need for subsurface heterogeneity characterization.

Over the past two decades, hydraulic tomography (HT) has been proposed (e.g., *Gottlieb and Dietrich*, 1995; *Yeh and Liu*, 2000) as a new method to characterize the heterogeneous distributions of aquifer parameters. During hydraulic tomography experiments, water is pumped or injected sequentially at various locations of the aquifer and pressure head changes at different locations and elevations are observed. The extensive data col-

lected from HT tests are then interpreted through inverse modelling to yield heterogeneous K and/or Ss fields for the target area. *Yeh and Liu* (2000) developed the first inverse modelling approach, steady state hydraulic tomography (SSHT), and demonstrated that the usefulness of the method through computational experiments and a preliminary sandbox study (*Liu et al.*, 2002). Later, more extensive laboratory sandbox (*Illman et al.*, 2007, 2008) and field (*Berg and Illman*, 2013) validation experiments were carried out to show the robustness of this approach. SSHT uses the steady state pressure head data from the drawdown curves and estimates only the K distribution and its corresponding uncertainty. With laboratory experiments, *Illman et al.* (2010a) compared several traditional approaches (i.e., permeameter tests, kriging) to SSHT through the prediction of pumping tests not used during the calibration effort. Results showed the superiority of SSHT, which accurately characterized the distributions of low and high K distribution and provided better drawdown predictions of independent pumping tests.

Later, *Zhu and Yeh* (2005) extended the hydraulic tomography method to use transient drawdown curves to estimate the K and S_s tomograms simultaneously, and demonstrated this approach with one hypothetical three-dimensional heterogeneous aquifer. *Liu et al.* (2007) validated transient hydraulic tomography using pumping test data in one controlled sandbox with a prescribed heterogeneity pattern. *Berg and Illman* (2011a) then compared the results from transient hydraulic tomography to traditional methods of characterization using the sandbox data collected by *Illman et al.* (2010a). *Berg and Illman* (2011b) first assessed this approach through field work at one field research site at the University of Waterloo and concluded that the inverse estimation of K and S_s using multiple pumping test data performed better than using data from single pumping tests.

As transient hydraulic tomography is computationally intensive, *Zhu and Yeh* (2006)

developed a THT method that utilizes the zeroth and first temporal moments of drawdown recovery data, instead of directly using the drawdown-recovery data. This approach was then tested in one sandbox experiment by *Yin and Illman* (2009). The results of estimated K tomogram was found to be comparable to the one estimated through THT by *Liu et al.* (2007) using pumping tests data from the same sandbox aquifer. However the S_s tomogram was not as robust, perhaps due to the loss of information on heterogeneity from the drawdown-recovery curve resulting from the smoothing nature of the temporal moment approach.

Along with the development of HT method, studies on hydraulic tomography have been conducted through computational experiments (*Bohling et al.*, 2002; *Brauchler et al.*, 2007; *Xiang et al.*, 2009; *Castagna and Bellin*, 2009; *Cardiff et al.*, 2009; *Cardiff and Barrash*, 2011; *Liu and Kitanidis*, 2011; *Schöniger et al.*, 2012), controlled sandbox experiments (*Liu et al.*, 2002; *Brauchler et al.*, 2003; *Liu et al.*, 2007; *Illman et al.*, 2007, 2008, 2010a,b, 2015; *Yin and Illman*, 2009; *Berg and Illman*, 2011a; *Liu and Kitanidis*, 2011; *Schöniger et al.*, 2015; *Zhou et al.*, 2016) and field pumping tests (*Bohling et al.*, 2007; *Illman et al.*, 2009; *Brauchler et al.*, 2010, 2011, 2012; *Castagna et al.*, 2011; *Huang et al.*, 2011; *Berg and Illman*, 2011a, 2013; *Cardiff et al.*, 2009, 2012, 2013a; *Berg and Illman*, 2015).

Among the various studies, steady state (*Illman et al.*, 2007; *Cardiff et al.*, 2009), temporal moment (*Zhu and Yeh*, 2006; *Yin and Illman*, 2009), transient (*Zhu and Yeh*, 2005; *Liu et al.*, 2007), travel time (*Brauchler et al.*, 2011), and oscillatory pressure signal (*Cardiff et al.*, 2013b) data have been inverted to map heterogeneity patterns. Compared to hydraulic parameters estimated via traditional approaches, such as kriging of local scale data or effective parameters by treating the aquifer to be homogeneous, *Illman et al.* (2010a) and *Berg and Illman* (2011a, 2015) found that hydraulic parameters estimated

through HT are significantly better in predicting drawdown responses. Furthermore, *Ni et al.* (2009) through synthetic simulations and *Illman et al.* (2012) through sandbox experiments demonstrated that HT results can significantly improve the predictions of solute transport. *Sun et al.* (2013) conducted numerical experiments to investigate sampling time strategy and boundary effects on the HT estimations. Although the extensive body of research shows promising results of HT, there are still further aspects that need to be investigated, for both laboratory works and potential applications of HT in the field. Thus, this thesis focuses on answering the following questions:

- Can hydraulic tomography be applied in unconfined aquifers where unsaturated flow processes are important and cannot be ignored?
- How useful is geological information for hydraulic tomography analysis under laboratory and field experimental conditions?

1.2 Geostatistical Inversion Approach

In this thesis, the simultaneous sucessive linear estimator (SimSLE) developed by *Mao et al.* (2013c) for interpreting sequential pumping tests in unconfined sandbox aquifers was used for both sandbox and field studies. This estimator is built from the successive linear estimator (SLE) by *Yeh et al.* (1996) and an earlier version of SimSLE by *Xiang et al.* (2009) for jointly interpreting sequential pumping tests, but has extended functions to analyze variable saturated flow data. The SimSLE algorithm has a built-in finite element code MMOC3 (for modified method of characteristics) (*Yeh et al.*, 1993) to simulate groundwater flow and solute transport under variably saturated conditions.

In this algorithm, groundwater flow in an unconfined aquifer was assumed to satisfy the Richard's equation, which takes the following form:

$$\nabla \cdot [K(\psi) + \nabla(\psi + z)] + Q = [\beta \cdot S_s + C(\psi)]\frac{\partial \psi}{\partial t} \tag{1.1}$$

For steady state flow in a fully saturated heterogeneous aquifer, the equation could be simplified as:

$$\nabla \cdot [K\nabla(\psi + z)] + Q = 0 \tag{1.2}$$

subject to boundary and initial conditions:

$$\psi|_{\Gamma_1} = \psi_1, \quad -K(x)\nabla(\psi + z)|_{\Gamma_2} = q, \quad \psi|_{t_0} = \psi_0 \tag{1.3}$$

where, ∇ is the gradient operator, $K(\psi)$ is unsaturated hydraulic conductivity term, which equals to saturated hydraulic conductivity K when $\psi \geq 0$, z is the elevation head, Q is the source/sink term, β is a saturation index, which equals 1 if $\psi \geq 0$ and equals 0 if $\psi < 0$, S_s is specific storage, $C(\psi)$ is the specific moisture capacity, and t is time. ψ_1 is constant head at boundary Γ_1, q is the specific flux at boundary Γ_2, and ψ_0 is initial pressure head at time 0.

The model developed by (*Gardner*, 1958) is used to represent the relationship between $K(\psi)$ and ψ:

$$K(\psi) = K \cdot e^{(\alpha\psi)} \tag{1.4}$$

where α is a soil parameter representing the decreasing rate of K induced by decreasing ψ. The corresponding $\theta - \psi$ relationship developed by (*Russo*, 1988) is the following:

$$\theta(\psi) = \theta_r + (\theta_s - \theta_r) \cdot [e^{0.5\alpha\psi}[1 - 0.5\alpha\psi]]^{\frac{2}{2+m}} \tag{1.5}$$

where θ_s is the saturated water content, θ_r is the residual water content, m is a soil tortuosity related parameter, and m is set to zero.

The flow equation is solved using the Galerkin finite element technique with either the Picard or the Newton-Raphson iteration scheme in MMOC3 (*Yeh et al.*, 1993).

Parameter estimation

SimSLE is a cokriging like inversion estimator and it conceptualizes the spatially varying natural log values of a hydraulic parameter (e.g., LnK, LnS_s) as a random field. The parameter field is first cokriged by conditioning on available hard data (e.g., measurements of K and/or the observed pressure heads). During this step, the unconditional mean values, variances, and correlation scales are needed as initial guesses and the covariance function of the parameter should be known. In all the studies included in this thesis, the exponential model is used for the covariance functions of the parameter fields. The cokriged parameter field is then used in MMOC3 (*Yeh et al.*, 1993) to solve the flow equation to obtain the simulated pressure heads. Then, an iterative procedure is employed by SimSLE to successively minimize the differences between simulated and observed pressure heads. The iterative process continues until the difference between the two continuous estimated hydraulic parameter fields or the largest head difference between simulated and observed is smaller than a specified tolerance. More details to the algrorithm can be found in (*Mao et al.*, 2013c) and (*Xiang et al.*, 2009).

1.3 Structure of the thesis

This thesis contains results obtained from three studies. To clearly present each study, I have summarized the objectives and works in Chapter 2 for each published paper. Results of three papers are separately shown and discussed in Chapter 3, 4 and 5. Chapter 6 draws conclusions from each study and briefly gives some recommendations for potential research in the future.

Objectives and Contributions

HT studies have been studied extensively using numerical models (e.g., *Yeh and Liu*, 2000; *Liu et al.*, 2002; *Bohling et al.*, 2002; *Zhu and Yeh*, 2005; *Xiang et al.*, 2009; *Cardiff et al.*, 2009; *Cardiff and Barrash*, 2011), controlled sandbox experiments (e.g., *Liu et al.*, 2007; *Illman et al.*, 2007, 2008, 2010a; *Yin and Illman*, 2009; *Berg and Illman*, 2011a) and field pumping tests (e.g., *Bohling et al.*, 2007; *Illman et al.*, 2009; *Huang et al.*, 2011; *Brauchler et al.*, 2011; *Berg and Illman*, 2011b, 2013, 2015; *Cardiff et al.*, 2009, 2012, 2013a). To date, only a few studies describing the performance of HT tests in unconfined aquifers have been conducted (*Cardiff et al.*, 2009; *Cardiff and Barrash*, 2011; *Zhu et al.*, 2011; *Berg and Illman*, 2012; *Mao et al.*, 2013c).

Cardiff et al. (2009) presented a study using a potential-based HT inversion approach in an unconfined aquifer at the Boise Hydrogeophysical Research Site (BHRS), Idaho. They only estimated the K of the unconfined aquifer. Through numerical experiments, *Cardiff and Barrash* (2011) investigated the possibility of mapping the heterogeneous K and homogeneous storage terms using transient hydraulic tomography (THT). Unsaturated flow was ignored and instantaneous drainage of water was assumed from the unsaturated zone as the water table drops.

Field studies showed that ignoring flow in the unsaturated zone and assuming gravity drainage of water due to falling of the water table generally leads to estimated specific yield

values that are substantially smaller those that would be expected on the basis of other methods of measurement (see *Nwankwor et al.*, 1984; *Endres et al.*, 2007). *Mao et al.* (2011) and *Yeh et al.* (2012) emphasized that the traditional analyses of pumping tests in unconfined aquifers (e.g., delayed yield, (*Boulton*, 1954, 1963), and delayed water table responses, *Neuman* (1972)) yield significantly smaller specific yield values (e,g,. *Nwankwor et al.*, 1984; *Endres et al.*, 2007) and are physically incorrect. They suggested that a variably saturated flow based mathematical model, which considers more realistic water release mechanisms, is more appropriate for the analysis of drawdowns due to pumping in unconfined aquifers.

Through a sandbox study, *Berg and Illman* (2012) concluded that considering the heterogeneity of saturated zone parameters and that using an accurate effective value of the unsaturated parameters is sufficient to accurately predict the drawdown response in the unconfined aquifer. The numerical experiment study by *Mao et al.* (2013c) showed the potential of characterizing unconfined aquifers using HT data while considering variably saturated processes. To date, the joint estimation of both saturated and unsaturated zone parameters using the HT approach have not been demonstrated through laboratory or field experiments.

On the other hand, in a wide range of applications involving geological modelling, geological data available at low cost usually consist of documents such as cross-sections or geological maps and data from borehole logs or outcrop descriptions. Usually, the geological information available is very helpful to build site specific geological models. Rarely, these conceptualized models are calibrated to multiple pumping test datasets to estimate the hydraulic parameters. *Bohling et al.* (2003) calibrated a simplified geological model with equal-thickness layers as well as zonation based radar profiling. Results show that radar-based zonation provide remarkably good correspondence with the direct-push slug test

profiling, suggesting the importance of accurately delineating the layering. Compared to the highly parameterized model adopted by a geostatistical inversion approach to interpret the multiple and extensive pumping test data during HT tests, geological models provide simpler ways to conceptualize the aquifers. When the hydraulic parameters are assumed to be constant within the same layer, the inversion effort is considerably reduced due to the fewer number of parameters that need to be estimated compared to geostatistical inverse approaches which suffer from the non-uniqueness issue due to the highly parameterized nature of the approach and insufficient observation data.

HT based on geostatistical inverse methods usually begins with a homogeneous initial guess, which has been shown to be quite suitable when the pumping test data is abundant (e.g., *Yeh and Liu*, 2000; *Liu et al.*, 2002; *Zhu and Yeh*, 2005). Under field conditions, the pressure head data alone might not be enough to perfectly characterize all important features of the aquifer with high heterogeneity and when data densities are low. For example, HT inversion results at the *NCRS* revealed a high K zone for the bottom clay layers (*Berg et al.*, 2011), indicating that more pressure head data or additional types of information (e.g., geology or flux measurements) might be helpful in order to better characterize the bottom clayey zone. Specially, geological data, as well as the seismic or ground penetrating radar data could be quite abundant that could provide structural information of subsurface units. In this thesis, I focus on including of geological data for HT analysis.

Thus, the main objectives of this thesis are:

- Validation of hydraulic tomography in a laboratory sandbox unconfined aquifer that considers variably saturated flow processes;
- Investigating the value of geological information for HT interpretation through laboratory sandbox and field experiments.

Part I: Hydraulic Tomography in Unconfined Aquifer: Sandbox Study

This part of the thesis addresses the scientific question that whether it is possible to implement the HT survey and SLE algorithm that considers the variably saturated flow for estimating heterogeneous saturated and unsaturated properties with real data. This concept is tested in one controlled sandbox aquifer constructed by *Berg and Illman* (2012) and the same data are used for this study. Details to this work are published in a paper titled, "Validation of hydraulic tomography in an unconfined aquifer: A controlled sandbox study" by *Zhao et al.* (2015).

Specifically, two heterogeneous model cases are designed and compared: (1) Case 1, K, S_s, α and θ_s fields are estimated; (2) Case 2, only K, S_s are estimated and the unsaturated zone parameters for α and θ_s are assumed to be homogeneous. For both cases, I selected five pumping test data from a HT survey conducted in a laboratory unconfined aquifer and tested its ability to characterize the heterogeneity in saturated and unsaturated zone hydraulic parameters. Then I validated the estimated tomograms of K, S_s, and unsaturated zone parameters [pore size parameter (α and saturated water content (θ_s)] for the Gardner-Russo's model to predict pumping tests not used during model calibration and the results are compared to those based on the homogeneous assumption.

Part II: Geological model and Hydraulic tomography: Sandbox Study

In this part, the importance of geological data in Hydraulic Tomography (HT) was investigated through sandbox experiments. The pumping test data used in this work was obtained from the same sandbox as *Illman et al.* (2010a). In the work of *Zhao et al.* (2016), geological data was collected along boreholes for the interpretation of HT surveys. Four groundwater models with homogeneous geological units constructed with borehole data of varying accuracy are jointly calibrated with multiple pumping test data of two different pumping and observation densities. Model calibration and validation performances are quantitatively assessed using drawdown scatterplots and the results are compared to those from a geostatistical inverse model. Thus, this part of thesis systematically investigates the value of geological data for the interpretation of HT surveys.

In total, steady state head data from 24 cross-hole pumping tests were selected for model calibration and validation purposes. Two study cases were designed using different numbers of pumping and observation data for the purpose of model calibration and validation. Case 1 used pressure head data from eight pumping tests and 47 observation ports for model calibration. In Case 2, I chose four pumping tests and 15 observation ports for calibration , representing the situation with limited amount of data available.

On the other hand, four uncalibrated geological models were populated with permeameter K values of each sand type and used as initial values for SimSLE, shown as Case 3 in *Zhao et al.* (2016). In Case 3, the pumping test data used as well as other initial inputs for SimSLE were identical to Case 2 in terms of the number of pumping tests and monitoring points. Thus, this case enabled the examination of the potential impact of using geological models constructed from borehole data as prior distributions for geostatistical inversions.

Part III: Geological model and Hydraulic tomography: NCRS Field Study

In this part, I continue the work of *Zhao et al.* (2016) on evaluating the importance of geological data for HT analysis, using data collected from a highly heterogeneous glaciofluvial aquifer-aquitard sequence at a well-characterized field site. Model calibration and validation performances are compared among three approaches: (1) the effective value approach by treating the site to be homogeneous; (2) two geological zonation approaches treating each layer to be homogeneous; and (3) the highly parameterized geostatistics approach using the Simultaneous Successive Linear Estimator (SimSLE) code developed by *Xiang et al.* (2009).

The experimental data was collected at the North Campus Research Site (NCRS) located on the University of Waterloo Campus, in Waterloo, Ontario, Canada. Previous Quaternary geology investigations showed that main deposits in the study area belong to the Waterloo Moraine (*Farvolden et al.*, 1987; *Karrow*, 1979, 1993; *Sebol*, 2000), which consists of multiple glacial till layers resulting from the advance and retreat of glaciers. The continuous core samples obtained at the NCRS revealed sequentially deposited tills, from younger to older age, as the Tavistock Till, Maryhill Till and Catfish Creek Till (*Karrow*, 1979; *Sebol*, 2000).

In the study area, the surface till is recognized as the Maryhill Till, composed mainly of silty clay and accompanied with few stones (*Karrow*, 1979). The younger Tavistock Till only exists rarely and mainly as erosional remnants. Underlying the Maryhill Till is the Catfish Creek Till, which consists of stiff stony silt to sandy silt and is considered to be extremely dense. This till is commonly hard and difficult to drill, and is referred as the general base for our study.

A total of 14 pumping tests at the NCRS were selected for the HT data analysis. Nine

pumping tests were initially conducted by *Berg and Illman* (2011b). Additional pumping and injection tests were conducted during the fall seasons of 2013 and 2014 to obtain more complete hydraulic response dataset for the site. Only the late time pressure heads from ports indicating steady or quasi-steady state are selected for HT analysis. Among them, seven pumping tests are used for calibration, while the other seven pumping tests are selected for model validation purposes.

Two cases are considered (Case 1a and Case 1b) in the effective parameter approach. Case 1a treats the aquifer to be isotropic, where only K_{eff} is estimate and Case 1b treats the entire simulation domain to be anisotropic, for which we estimate the effective K_x, K_y and K_z.

Borehole logs of 18 pumping and observation wells are summarized from previous work (*Sebol*, 2000; *Alexander et al.*, 2011) and complied for the construction of the geological model. Based on the soil types and corresponding depth information, 19 different layers representing seven different material types are defined along all boreholes. The layer information between boreholes at different locations are interpolated to construct a three-dimensional geological model with dimensions of $70m \times 70m \times 17m$. Based on the layering and soil types, two geological models (Case 2a and 2b) with different numbers of layers are prepared for model calibration. One model consists of five layers (Case 2a), while the other model includes all 19 layers (Case 2b). The five-layer geological model is constructed by merging some layers with similar material, specifically layers 1 through 10 as layer 1*, layers 12 through 14 as layer 12*, and layers 16 through 19 as layer 16*, which are mainly composed of relatively low K clay and silt. Layers 11 and 15 are treated as two separate zones for the highly permeable nature of sand or sand and gravel. The five-layer model is constructed as a simplified model that only generally reflects the main high and low

permeable zone features, while the 19-layer geological model is used to take full advantage of the interpolated stratigraphy information. Additionally, once the geological model is constructed, no adjustments are made to the geological structure during the calibration phase.

Four scenarios (Case 3a, 3b, 3c and 3d) are considered for the geostatistical inversion approach using different prior distributions. For Case 3a, a uniform mean K field is used as the prior distribution for the geostatistical inversion before SimSLE starts to iteratively condition the parameter field with pressure head measurements. For the other three cases (Cases 3b – 3d), geologic information is used as prior knowledge for the inversion. Specifically, Case 3b use the estimated K values from Case 2a as the prior distribution; Case 3c use the K estimates from Case 2b as the prior distribution; Case 3d use the 19-layer geological model (Case 2b) populated with permeameter tested K values as the prior distribution.

Hydraulic Tomography in Unconfined Aquifer: Sandbox Study

The pumping test data used for this study are obtained by *Berg and Illman* (2012) in one controlled sandbox aquifer. Nine pumping tests (Ports 3, 12, 13, 14, 31, 35, 37, 39 and 43) conducted within the laboratory sandbox unconfined aquifer (Figure 3.1) have been selected for HT analysis. Five tests (Ports 3, 12, 14, 31 and 43) are used for geostatistical inverse modeling, and the other four tests (Ports 13, 35, 37 and 39) are kept for model validation. Prior to the joint analysis of multiple pumping test data using SimSLE, saturated and unsaturated zone parameters are estimated to provide reasonable initial inputs for the geostatistical inversion. Then, I performed the very first validation work of THT in a laboratory unconfined sandbox aquifer through jointly estimating hydraulic parameters for both saturated zone and unsaturated zone.

3.1 Methods

3.1.1 Sandbox Description

The sandbox used for the validation of HT has been described in detail by *Berg and Illman* (2012). Here, we only provide basic information of the sandbox, various equipment, and additional work that has been completed more recently. The sandbox has dimensions of 244

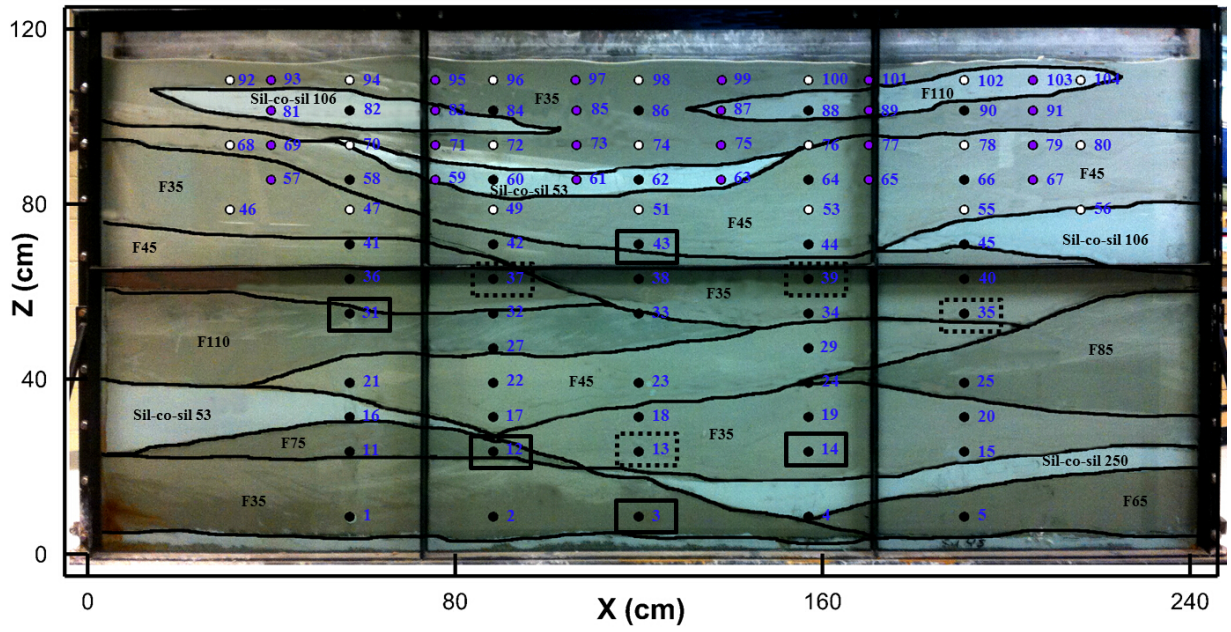


Figure 3.1: Front view of sandbox aquifer used for pumping tests showing layers and port locations. Solid black circles indicate the pressure transducer ports; solid white circles indicate the tensiometer ports; solid purple circles indicate the water content sensor ports. Solid squares indicate the pumped ports to generate the drawdown data for inverse modeling (calibration), while the dashed squares are pumped locations to generate data for validation purposes.

cm in length, 122 cm in height, and 9.4 cm in depth. It was filled with different sediments to construct a heterogeneous unconfined aquifer consisting of 17 layers. In order to simulate unconfined aquifer flow conditions, we kept the top of the aquifer open to the atmosphere, while there was no flow through all other sides of the sandbox. Pressure heads were collected with 47 pressure transducers (Model S35; BHL Instruments) and 22 column tensiometers (Model CL-029B; Soil Measurement Systems) installed within the upper zone experiencing unsaturated flow, while water content data was collected with 24 water content sensors (Model EC5; Decagon Devices Inc.). The tensiometers were equipped with Microswitch pressure transducers. Figure 3.1 shows the front view of the sandbox, showing the port and instrument locations as well as the layout of different layers.

Six different types of commercially sieved sands (F35, F45, F65, F75, F85 and F110) and four types of silts (Sil-col-sil 45, Sil-col-sil 53, Sil-col-sil 106 and Sil-col-sil 250) were

used to pack the sandbox. In particular, three lenses were packed at the top part of the sandbox in an attempt to simulate aquifer heterogeneity in the unsaturated zone during the pumping test. These three lenses, especially the silt lenses, have different saturated and unsaturated hydraulic properties compared to the surrounding sandy material. Thus, successfully mapping these lenses through HT may provide a visual test to the performance of the SLE approach.

3.1.2 Description of Pumping Tests

Nine pumping tests were conducted in a tomographic fashion to stress the sandbox aquifer at different locations (port 3, 12, 13, 14, 31, 35, 37, 39 and 43) at a constant pumping rate of 60 ml/min with a peristaltic pump. Prior to each pumping test, the water level was kept steady at an initial level of 112 cm from the bottom of sandbox to collect background hydraulic head levels. During each pumping test, we collected data every 0.25 seconds from the 46 pressure transducers and 22 tensiometers throughout all nine tests, which we found to be sufficient in capturing the rapid transient pressure change throughout the sandbox aquifer. The pumping test durations varied from approximately 1.8 h at port 43 located in the upper portion of the sandbox to nearly 7.5 h at port 3 located near the bottom of the sandbox. Among the nine pumping tests, data from five tests (ports 3, 12, 14, 31 and 43) were selected for HT analysis, while the other four test datasets were reserved for validation purposes.

3.1.3 Inverse Model Description

In this study, the THT analysis of pumping tests was carried out using the version of SLE developed by *Mao et al.* (2013c). Groundwater flow in an unconfined aquifer was assumed to satisfy the Richard's equation (1.1). Meanwhile, the Gardner-Russo's model developed by *Gardner* (1958) (equation 1.4) and *Russo* (1988) (equation 1.5) were used to represent the relationships between $K(\psi)$, θ and ψ for the unsaturated zone flow. In one early study by *Berg and Illman* (2012) for the same unconfined sandbox, the van Genuchten-Mulalem model (*van Genuchten*, 1980; *Mualem*, 1976) were found to be suitable to fit the observed water retention curves as well as suitable for the predictions of drawdown of different drainage tests. However, it has been shown that the high nonlinearity of the van Genuchten-Mulalem mdoel led to convergence issue for the current highly parameterized geostatistical inversion problem in the variable saturated flow condition. Thus, the Gardner-Russo's model was selected here for its computational simplicity instead of the widely used van Genuchten-Mulalem model.

3.1.4 Inverse Model Parameters

While the sandbox has overall dimensions of 244 cm in length, 9.4 cm in depth and 122 cm in height, the model domain is only 112 cm in height, reflecting the height of the packed sand and silt material within the sandbox. In order to characterize the hydraulic parameters, the sandbox aquifer was discretized into 3, 645 finite elements using variable element sizes, with an average dimension of 3 cm $\times$ 9.4 cm $\times$ 2.5 cm. This grid setting was kept the same with that was used in the previous THT analysis of the saturated zone of the same sandbox aquifer by *Berg and Illman* (2012). The top surface of the model domain is

assigned a no flow condition, since it is open to the atmosphere and no significant amount of water is gained or lost. The other five model boundaries (three sides and one in the front and another in the back) are treated as no flow boundaries. As we have no available hard data to condition the inverse model, only the pressure head data are used for the analysis. The initial model inputs needed for the SLE inversion are described as bellow (Case 1 for further discussion).

Initial values for $\boldsymbol{K}$, and S_s

The distribution of saturated properties had been characterized by *Berg and Illman* (2012) through a transient HT analysis of eight pumping tests under fully saturated conditions, which can be used to obtain estimates for saturated hydraulic properties. They also obtained effective values, $K_{eff} = 1.85 \times 10^{-2}$ cm/s, $Ss_{eff} = 3.94 \times 10^{-5}$ /cm, through calibrating the pumping test data at port 22. Since it is more likely to obtain effective parameters in the field, we use 1.85×10^{-2} cm/s and 3.94×10^{-5} /cm as the initial values.

On the other hand, to provide reasonable values for comparison of the saturated properties of different sandbox materials, K and S_s values were calculated for each material based on the HT tomograms in *Berg and Illman* (2012). Specifically, K and S_s values for the elements located in the same aquifer material, total of nine materials, were carefully selected based on the layout of aquifer material and geometric means of K and S_s were taken. The results are shown in Table 3.1.

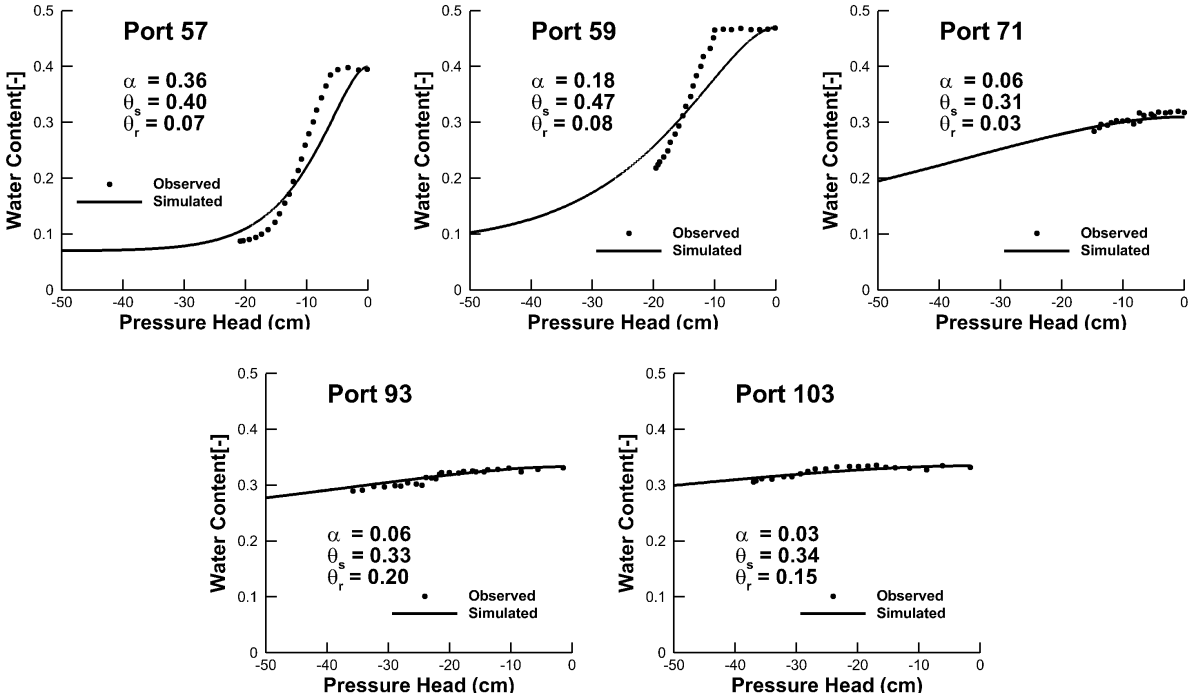


Figure 3.2: Moisture characteristic curves fit for different aquifer materials.

Initial values for α, θ_s, and θ_r

In order to obtain unsaturated hydraulic parameters for the upper areas of the sandbox aquifer which experience negative pressures, the aquifer was drained starting from fully saturated condition by pumping water at port 3. Meanwhile, the θ and ψ in the upper portion of the aquifer was recorded at the locations indicated on Figure 3.1. We utilized the parameter estimation program PEST (*Doherty*, 1994) to fit equation (1.5) to the observed θ and ψ data obtained from the drainage experiments, and estimated α with a 95% confidence interval for each material type. Results are shown in Table 3.1. Examples of the curve fittings for five ports located in different material are shown in Figure 3.2. As water content sensors were installed in the upper sandbox, we were only able to estimate the unsaturated zone parameters for the five sand types (F35, F45, F110, Sil-co-sil 53, Sil-co-sil 106). In Table 3.1, the numbers of fitting curves for each material type are also given, whereas N/A is used to indicate no available data for curve fitting. Geometric means of α, θ_s, and θ_r values

estimated at all water content ports from curve fitting of water content and closest pressure head data are calculated as: $\alpha = 1.3 \times 10^{-1}$ /cm, $\theta_s = 0.34$ and $\theta_r = 0.06$. These values are used in the inverse model as the initial guesses for calibration of five pumping test data.

Variance and correlation scales

The SLE algorithm requires initial estimates of variances (σ^2) and correlation scales (λ_x, λ_z) of all estimated parameters (ln K, ln S_s, ln α, and ln θ_s) for the inverse model. Usually, these statistical properties can be calculated based on available core samples. Through numerical experiments, *Yeh and Liu* (2000) concluded that these statistical parameters produce minor impacts on the inverse modeling results of K, especially when ample head measurements are available. Since it is impossible to get precise values for these statistical properties when the true distributions are unknown, we select the horizontal correlation scales of K and S_s as $\lambda_x = 150$cm used in *Berg and Illman* (2012), while using the vertical correlation scale $\lambda_z = 30$cm. Meanwhile, $\sigma^2_{lnK} = 1.0$ and $\sigma^2_{lnSs} = 0.5$ are used considering that estimated S_s tomograms show less heterogeneous distributions compared to K tomograms in previous studies (*Liu et al.*, 2007; *Xiang et al.*, 2009; *Berg and Illman*, 2011a, 2012). Since only unsaturated zone parameters in the upper portion of the sandbox are being estimated, and unsaturated zone heterogeneity only has minor impacts in forming the S-shaped drawdown curves (*Berg and Illman*, 2012), smaller correlation scales ($\lambda_x =$ 100cm, $\lambda_z = 10$cm) and smaller variances ($\sigma^2_{ln\alpha}$ =0.02 and $\sigma^2_{\theta_s}$ =0.001) are used as initial estimates for α and θ_s.

Table 3.1: Hydraulic Parameters Calculated for Each Material.

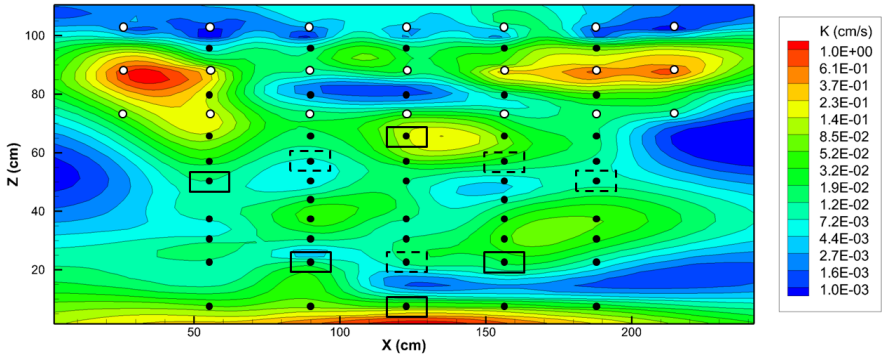
Material Type	$K(cm/s)^a$	$Ss(1/cm)^a$	θ_s^b	θ_r^b	$\alpha(/cm)^b$	Number of Curves
F35	2.63×10^{-2}	5.98×10^{-5}	0.34	0.04	0.20	6
F45	2.75×10^{-2}	4.84×10^{-5}	0.33	0.04	0.16	5
F65	5.60×10^{-3}	2.06×10^{-5}	N/A^c	N/A	N/A	N/A
F75	3.20×10^{-2}	1.65×10^{-5}	N/A	N/A	N/A	N/A
F85	8.78×10^{-3}	2.93×10^{-5}	N/A	N/A	N/A	N/A
F110	1.24×10^{-2}	3.89×10^{-5}	0.34	0.15	0.02	2
Sil-co-sil 53	2.38×10^{-2}	7.19×10^{-5}	0.38	0.09	0.13	6
Sil-co-sil 106	4.34×10^{-3}	5.65×10^{-5}	0.35	0.08	0.14	4
Sil-co-sil 250	4.38×10^{-3}	1.86×10^{-5}	N/A	N/A	N/A	N/A

[a] Data used for calculation of K and Ss for each material is obtained from *Berg and Illman* (2012).
[b] Data obtained through drainage experiment and fit to Gardner-Russo's Model.
[c] N/A represents that there are no data available for fitting Gardner-Russo's Model

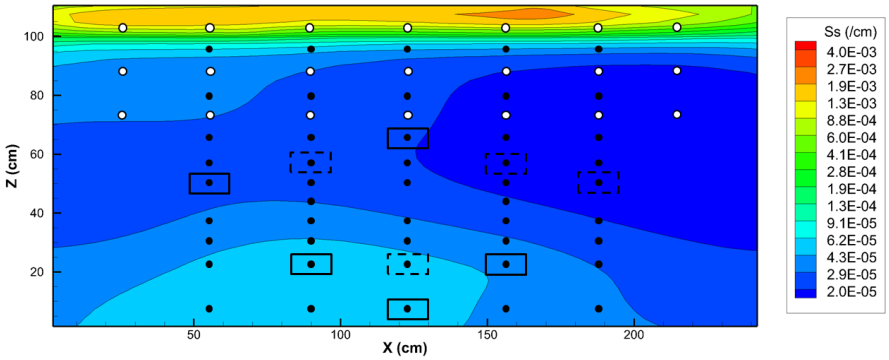
Data Selection

Based on the cross correlation analysis by *Mao et al.* (2013b), early time ψ data contain the most information on S_s heterogeneity, while the late time head data in the saturated zone contains the greatest information for K heterogeneity within the cone of depression, and ψ data in the unsaturated zone contain most information on α and θ_s. Thus, we selected ψ data from early, intermediate and late time based on the time-drawdown curves for calibration purposes. We note that pressure transducers installed in the sandbox can only record positive pressures. Therefore, the time-drawdown curves from upper portion ports 58, 60, 62, 64, 66, 82, 84, 86, 88, 90 show no additional drawdown after the water level drops below those ports. For these curves, only the early time ψ data are included in the HT analysis. In total, we used 480 data points from five different tests for the inverse modeling effort, including 400 data in the saturated zone and 80 data in the unsaturated zone. The HT analysis is conducted by inverting the data from five pumping tests simultaneously.

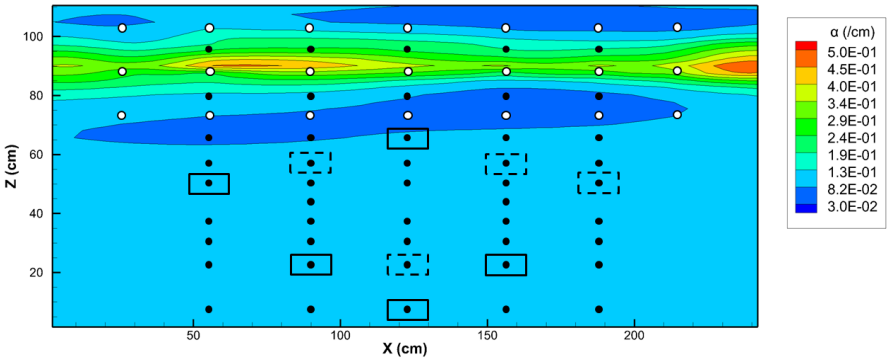
(a) K



(b) S_s



(c) α



(d) θ_s

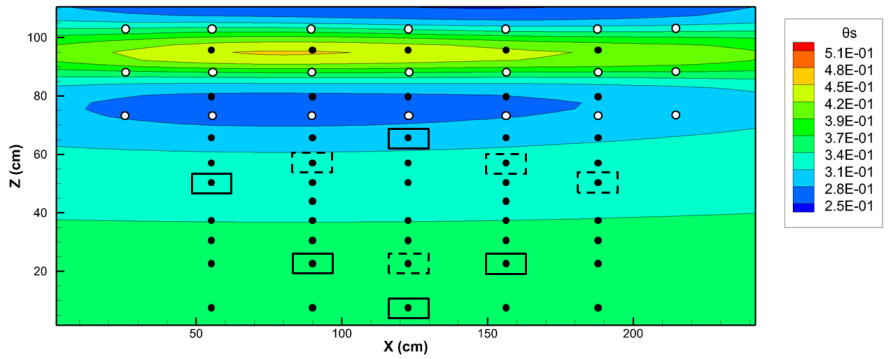


Figure 3.3: Estimated tomograms of a) K, b) S_s, c) α, d) θ_s using sandbox pumping test data. Solid squares indicate the pumping ports used for inverse estimation. Dashed squares indicate the pumping ports used for validation purposes. ∘ indicates locations for pressure transducers. • indicates locations for tensiometers.

Computational Costs

For the cost side of HT analysis, the inversion is executed with 32 processors on a PC-cluster consisting of 1 master and 2 slave nodes at the University of Waterloo. Each slave node has 16 processors running with an average RAM of 4GB. Calibration of five test transient data was completed with 64 iterations in two days using the SLE algorithm. Some initial estimates, like K and S_s described in section 3.1.4, are based on previous calibration of one single pumping test data in *Berg and Illman* (2012). Future application of HT in unconfined aquifers under field conditions may require some effort to obtain reasonable initial guesses when no previously collected data is available.

3.2 Results and Discussion

3.2.1 Visual comparison of $\boldsymbol{K}$, S_s, α, and θ_s tomograms with stratigraphy

The estimated tomograms for all four parameters (K, S_s, α, and θ_s) are shown in Figure 3.3, while Figure 3.5 shows the corresponding residual variance fields. The uncertainty in estimates is represented by the residual variance of each parameter, which is set as the value of initial estimate of parameter variance. After incorporating observed head data, the residual variance represents the updated residual cross-covariance between the observed heads and parameters to be estimated at the observed locations. Lower residual variance suggests higher confidence in the estimates.

Figure 3.3a shows the estimated K tomogram revealing that the major stratigraphic features shown on Figure 3.1 are captured quite accurately. To make this comparison

clearer, we superimpose the K tomogram Figure 3.3a over the stratigraphy Figure 3.1, as shown in Figure 3.4, which reveals that the low K zones are vividly captured. In particular, the low K zones in the upper and right hand side are the most obvious. The Sil-co-sil 53 layer and F110 layer in the left hand side also show a low K value, although it is not as pronounced.

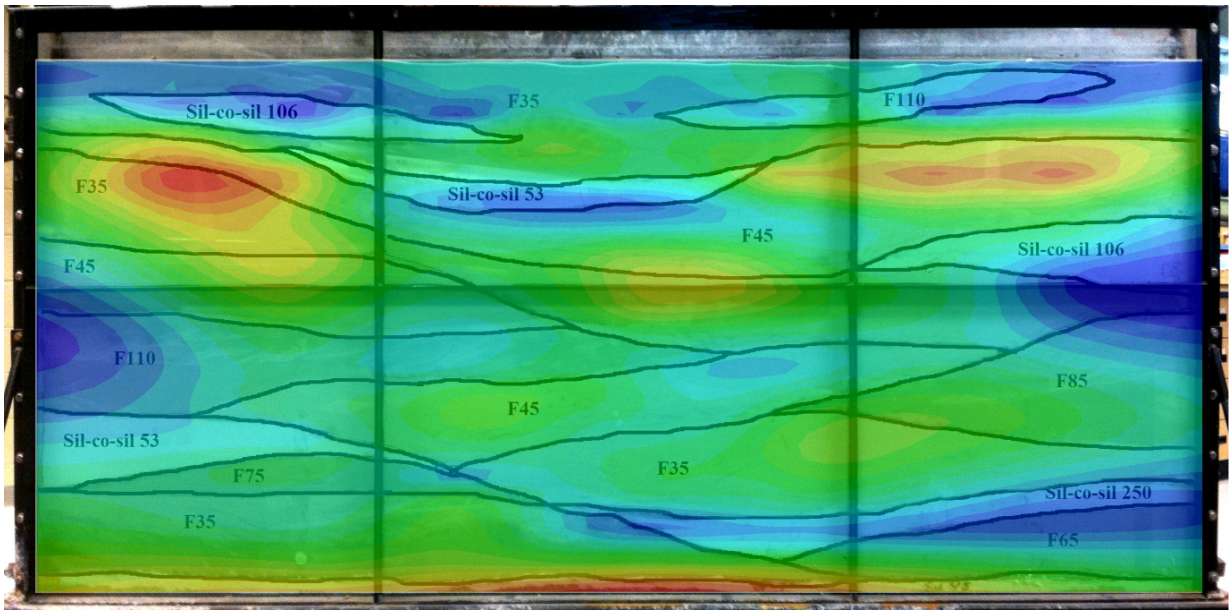


Figure 3.4: Overlay of estimated K tomogram over the sandbox aquifer.

Examination of Figure 3.5a shows that the residual variance is lower at measurement locations compared to the left and right boundaries regions. These results show that the joint inversion of five pumping test data is quite robust in capturing the high and low K patterns. This is critical for investigating contaminant transport problems in unconfined aquifers, where correctly locating the low K layers will be beneficial, as those units can store and release contaminants over long periods.

Compared to the K tomogram, the estimated S_s tomogram (Figure 3.3b) is considerably smoother and shows no obvious layering. Physically, the S_s term is related to the porous medium and water compressibility. The overall low S_s values, not including the zone along the top boundary, seem to suggest that the porous medium and water compressibility in

our unconfined sandbox aquifer is low in most of the model domain. This low S_s pattern is found to be suitable in simulating the early time aquifer responses shown in Figure 3.8 and Figures A1-A3 in the Supplementary Information section. Since the S_s term only is applicable in the saturated zone and is sensitive to drawdown curves at early times during a pumping test in an unconfined aquifer (*Zhu et al.*, 2011; *Mao et al.*, 2013b), the narrow and long region near the top boundary should not be considered to be reliable, although a low residual variance zone is shown in this narrow region (Figure 3.5b).

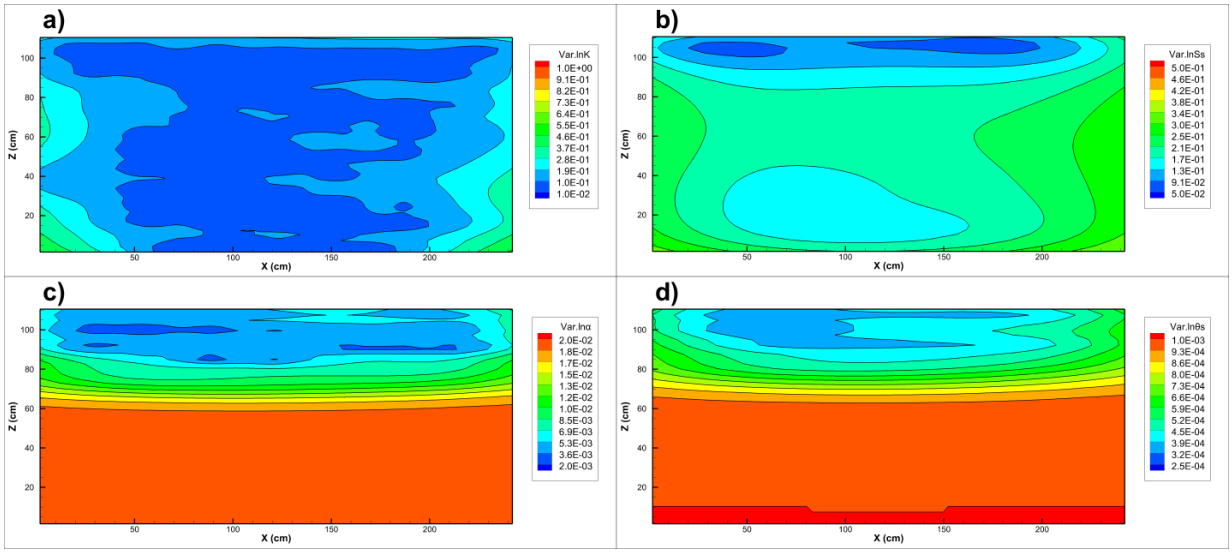


Figure 3.5: Residual variances of estimated tomograms for: a) ln K, b) ln S_s, c) ln α, d) ln θ_s.

For all pumping tests analyzed using the inverse model, only ψ data from the first 200 minutes is utilized and negative ψ are observed by tensiometers installed in the upper part of the sandbox. Thus, we anticipate the code to obtain reliable heterogeneous distributions of α and θ_s only in the upper region. Figures 3.3c and 3.3d show the α and θ_s tomograms, respectively. Although the match is not exact, generally speaking, the low values of the estimated α tomogram corresponds with the location of the low K layers, Soil-co-sil 106, Sil-co-sil 53 and F110, while a relatively high α value zone is assigned to the narrow region for F35 material in the upper part of the aquifer. This overall pattern is reasonable since

our estimates for α through fitting the water content and pressure head curves also gives high α values for high K sand material and low α for relatively low K material.

A similar layering pattern is also visible in the θ_s tomogram (Figure 3.3d). The relatively high K regions are estimated as high θ_s zone, while the low K regions are estimated as low θ_s zone. This may seem counterintuitive as low K material (i.e., silts and clays) typically have high θ_s values (which corresponds with high porosity), and also are different to the measured data by water content sensors, as shown in Table 3.1. This pattern change is possibly caused by slow and partial drainage of low K layers and the use of unified value of residual water contents (θ_r=0.06) for all the materials, whereas in reality the residual water content for each material is different. In the lower part of the sandbox, both α and θ_s show values close to the initial estimates, indicating that the inverse model does not change the tomograms significantly. This is expected because no negative pressure head data are available in this area of the sandbox and the inverse model is not sensitive to unsaturated flow parameters in the saturated zone which is reflected in the high residual variance values (Figures 3.5c and 3.5d). According to the above visual comparisons, we conclude that the HT analysis, using the variably saturated model and the SLE algorithm, is able to characterize both the saturated and unsaturated zone parameters quite reasonably using pumping test data from the unconfined aquifer in this sandbox.

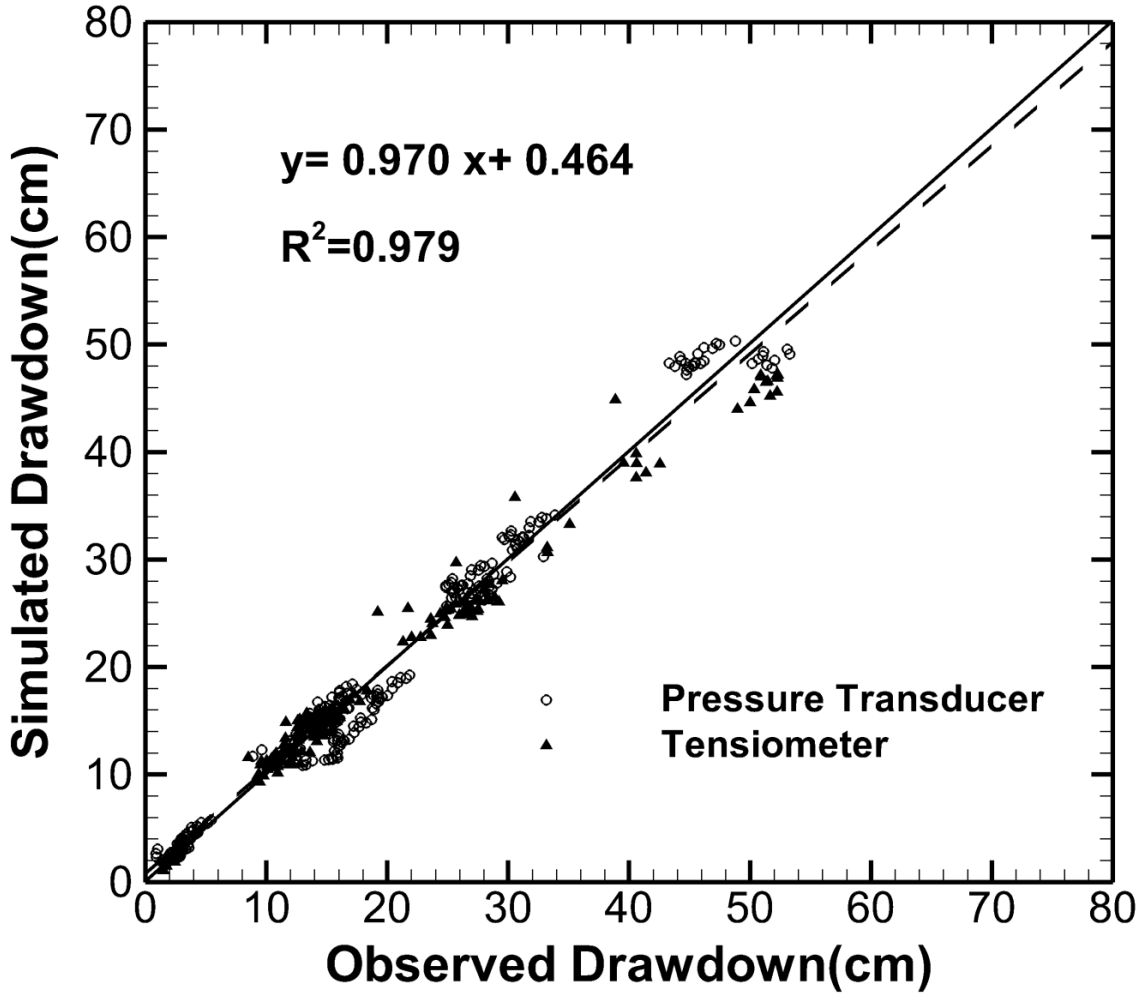


Figure 3.6: Overall calibration scatterplot. Black circles are data selected from pressure transducer ports, Solid triangles are data selected from tensiometer ports. Solid black line is 1:1 line, indicating a perfect match. Dash black line is the best fit. The linear fit results are also included.

3.2.2 Calibration and Validation Results

To evaluate the K, S_s, α, and θ_s tomograms generated by the THT analysis of the sandbox data, we examine the quality of model calibration and validate all four tomograms simultaneously through the forward simulation of the four pumping tests not used in inverse modeling effort.

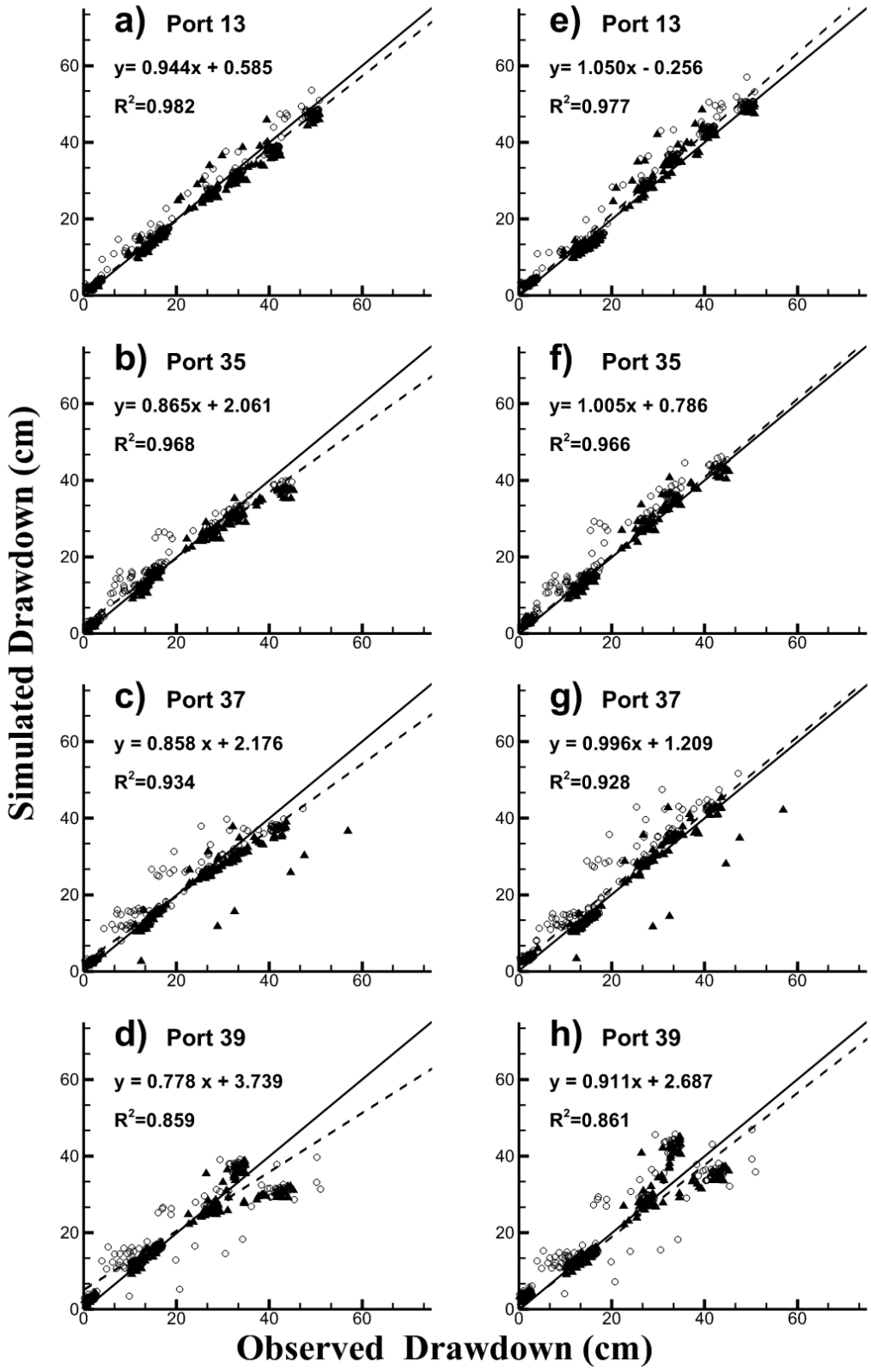


Figure 3.7: Validation Scatterplots for HT analysis Case 1 and Case 2 at different pumping locations. Results of Case 1 are shown in (a)-(d), while (e)-(f) show results for Case 2. The solid line is a 1:1 line indicating a perfect match. The dash line is a best fit line. The linear fit results are also included on each plot. Black circles are data selected from pressure transducer ports, solid triangles are data selected from tensiometer ports.

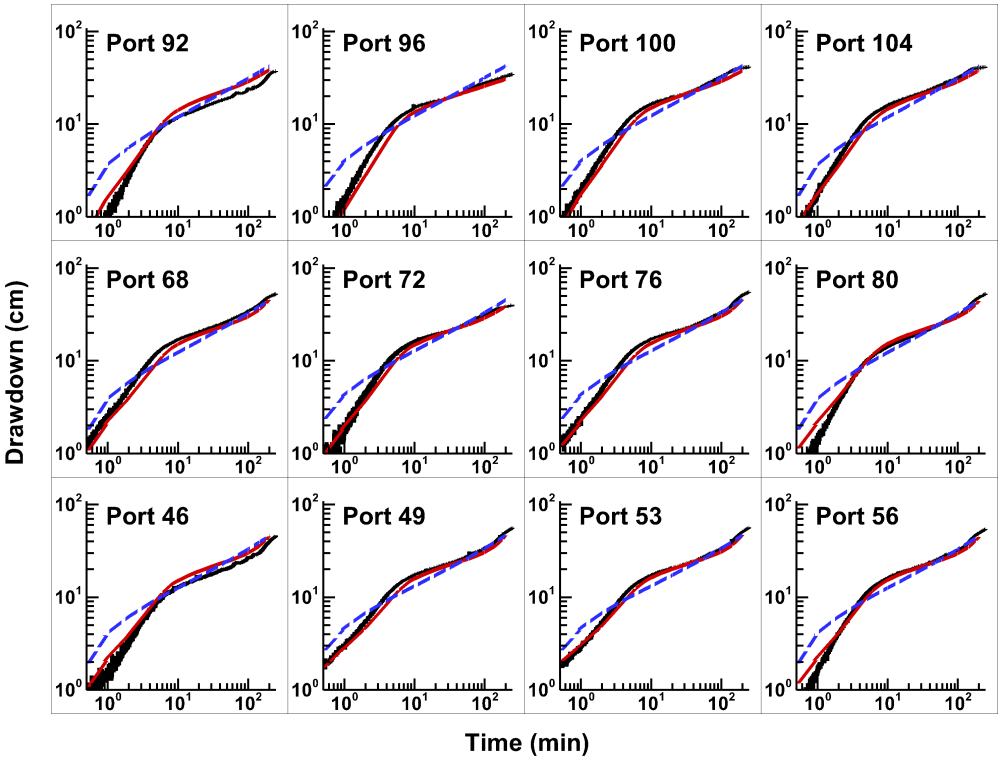
Model Calibration

We first examine the quality of model calibration by plotting the observed drawdown from the pressure transducers both in the saturated and unsaturated zones versus simulated drawdown using the estimated K, S_s, α, and θ_s tomograms. Figure 3.6 shows that the corresponding drawdowns are evenly distributed along the 1:1 line, and the coefficient of determination (R^2) between the simulated and observed values is 0.979, indicating a good match for both pressure transducers in the saturated and unsaturated zones.

Validation of Sandbox THT

Another more important way to evaluate the tomograms is to predict pumping tests that are not included in the HT analysis and to examine whether the various drawdown curves can be predicted accurately throughout the duration of the pumping test. Figure 3.7 shows the validation scatterplots for the pumping tests conducted at ports 13, 35, 37, and 39. In Figure 3.7a, drawdown data are selected at seven time points, t = 1, 6, 10, 60, 100, 150, 200 min, covering early, intermediate and late time stages. In Figure 3.7b, 3.7c, 3.7d, drawdown data at six time points, t = 1, 6, 10, 60, 100, and 150 min, are selected due to the shorter pumping durations at port 35, 37 and 39. The R^2 of all 4 tests vary from 0.859 at port 39 to 0.982 at port 13, indicating a fairly good correspondence between simulated and observed drawdowns.

(a) Tensiometer Ports



(b) Pressure transducer ports

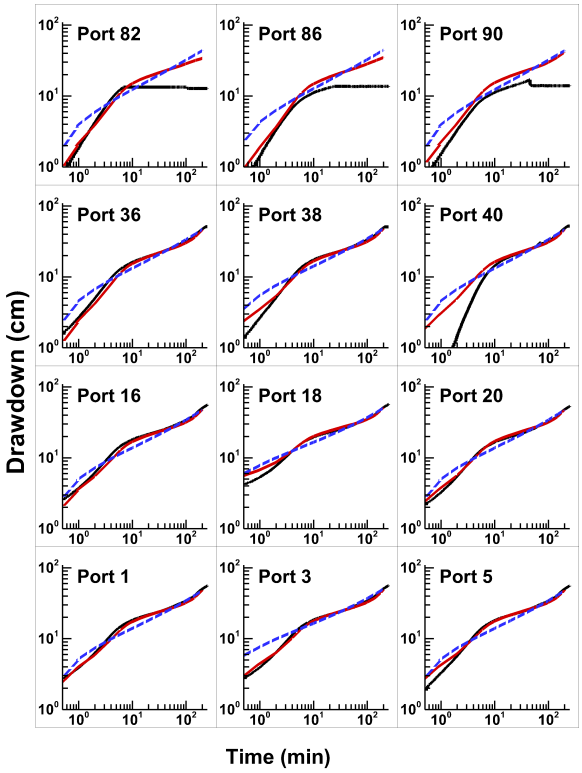


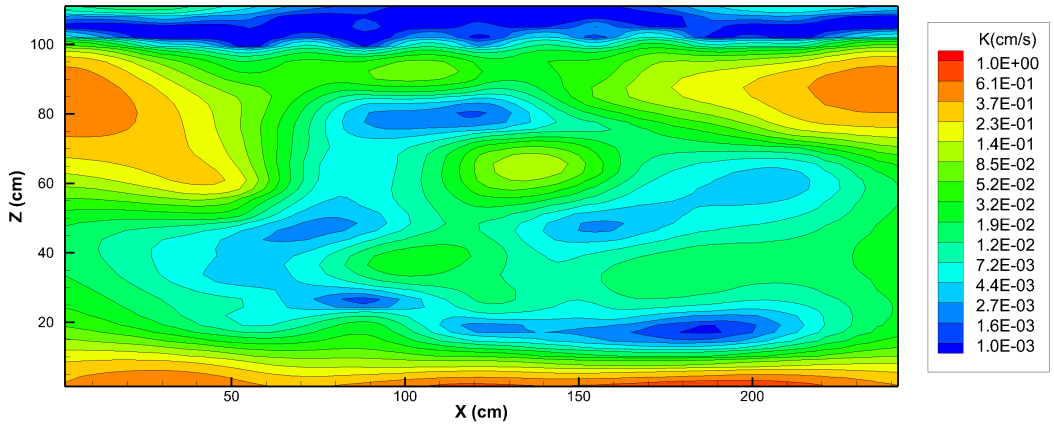
Figure 3.8: Simulated and observed drawdown at a) pressure transducer ports and b) tensiometer ports during the pumping test at port 13. The black solid lines are observed data; the red lines are simulated drawdown using results from HT; the blue dashed lines are simulated drawdown using homogeneous assumption. The layout of all the ports is identical to the true layout in the sandbox.

We also plot the simulated time drawdown curves for both pressure transducer ports and tensiometer ports against the observed time-drawdown curves for the pumping test at port 13 used for validation on Figure 3.8. The Supplementary Information section includes additional matches from ports 35, 37, and 39 as Figures A1 through A3, respectively. Examination of Figure 3.8 and Figures A1 – A3 in the Supplementary Information section reveals that, although there are slight misfits for one or two pressure transducer ports at the top of the sandbox, the overall match for both pressure transducer and tensiometer data is quite good, especially for the pumping tests at port 13 (Figure 3.8), port 35 (Figure A1) and port 37 (Figure A2). Here, we need to clarify that, since HT provides effective parameter fields conditioned on the 480 given data points, we cannot expect the simulated curves to perfectly match all the observed data. Based on these validation figures, we conclude that the joint inversion of multiple pumping tests can estimate the unknown saturated and unsaturated parameter fields quite well, which in turn benefit our ability to predict the pumping response at unconfined aquifers.

Effect of Unsaturated Zone Heterogeneity

We next investigated the effect of unsaturated zone heterogeneity on both estimating saturated zone parameters and subsequent drawdown prediction, by using fixed homogeneous unsaturated zone parameters $\alpha = 1.3 \times 10^{-1}$ /cm, $\theta_s = 0.34$ and $\theta_r = 0.06$ (Case 2 for the following discussion), while the other inputs are identical to Case 1 for SLE inversion. Results show that the inverted K and S_s tomograms (Figure 3.9a and 3.9b) have larger spatial variances ($\sigma^2_{lnK} = 2.92$, $\sigma^2_{lnSs} = 1.49$) compared to Case 1 ($\sigma^2_{lnK} = 2.30$, $\sigma^2_{lnSs} = 1.26$). This kind of increase was attributed to the assumption of uniform spatial distribution of unsaturated zone parameters in *Mao et al.* (2013c).

(a) K



(b) S_s

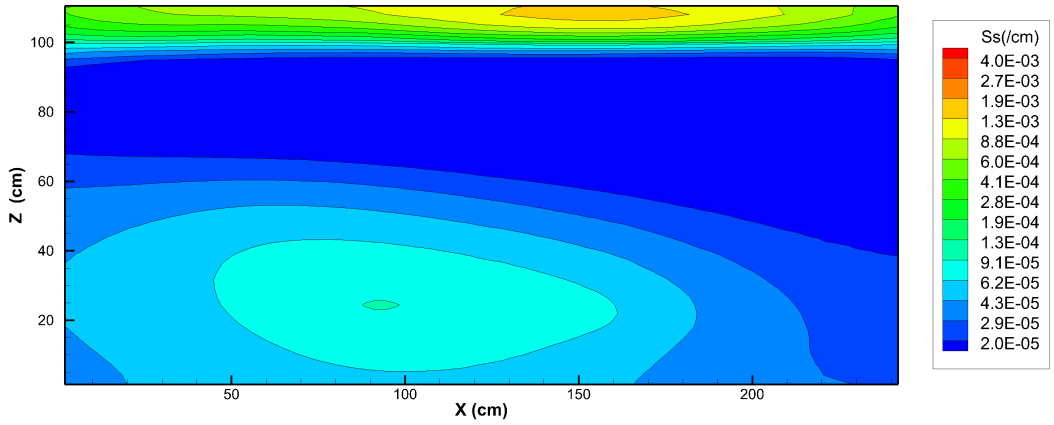


Figure 3.9: Estimated tomograms of a) K, b) S_s for Case 2, with fixed unsaturated parameter $\alpha = 1.3 \times 10^{-1}$/cm, $\theta_s = 0.34$ and $\theta_r = 0.06$. Spatial variances of the estimated tomograms are $\sigma^2_{lnK} = 2.92$, $\sigma^2_{lnSs} = 1.49$.

Overall, Figure 3.9a is relatively different from Figure 3.3a for K tomogram, while Figure 3.9b shows a similar S_s distribution pattern as Figure 3.3b. The K tomogram captures low value zones in both saturated and unsaturated zones of the central sandbox area. However, there is some loss in details with respect to the two low K zones near the top boundary. Through numerical experiments, *Mao et al.* (2013c) concluded that effective unsaturated parameters are sufficient for the estimation of saturated parameter patterns. The relatively apparent changes in Figure 3.9a compared to Figure 3.3a indicate that the impact of unsaturated zone is significant for identification of the layering pattern in our

sandbox study (also see Figure A4 for scatterplots), where the real drainage data was used instead of model-based numerical experiment data and a large extent of the unsaturated zone, that extended about 46 cm below initial water level (41.1% of the entire sandbox).

Figures 3.7e – 3.7h show the validation scatterplots of the same four pumping tests using the estimated parameters from Case 2. Note that the time points selected for simulated drawdowns are exactly the same to Case 1 as introduced earlier in Section 3.1.4. Compared to Figures 3.7a – 3.7d from Case 1, Figures 3.7e – 3.7h show quite close results for Case 2, indicating that the use of homogeneous unsaturated zone parameters can also satisfy drawdown prediction.

Comparison of HT Results with Homogeneous Model Results

We next simulate the drawdown response by assuming that all aquifer and unsaturated zone parameters are homogeneous. The effective hydraulic parameters of the sandbox material were estimated through coupling variably saturated flow model MMOC3 (*Yeh et al.*, 1993) with parameter estimation code PEST (*Doherty*, 1994). In order to provide a fair comparison, we simultaneously calibrated the 480 data selected from 5 pumping tests which are the same for SLE calibration. Such a large number of non-redundant data sets ensure the requirement of necessary conditions as outlined by *Mao et al.* (2013a) for the inverse problem for homogeneous aquifers to be well-defined. The effective parameters were estimated as $K_{eff} = 1.47 \times 10^{-2}$ cm/s and Ss_{eff}= 3.04×10^{-5} /cm for the saturated zone, and $\alpha_{eff} = 1.1 \times 10^{-1}$ /cm, $\theta_{seff} = 0.40$ for unsaturated zone, while fixing the residual water content, $\theta_r = 0.06$.

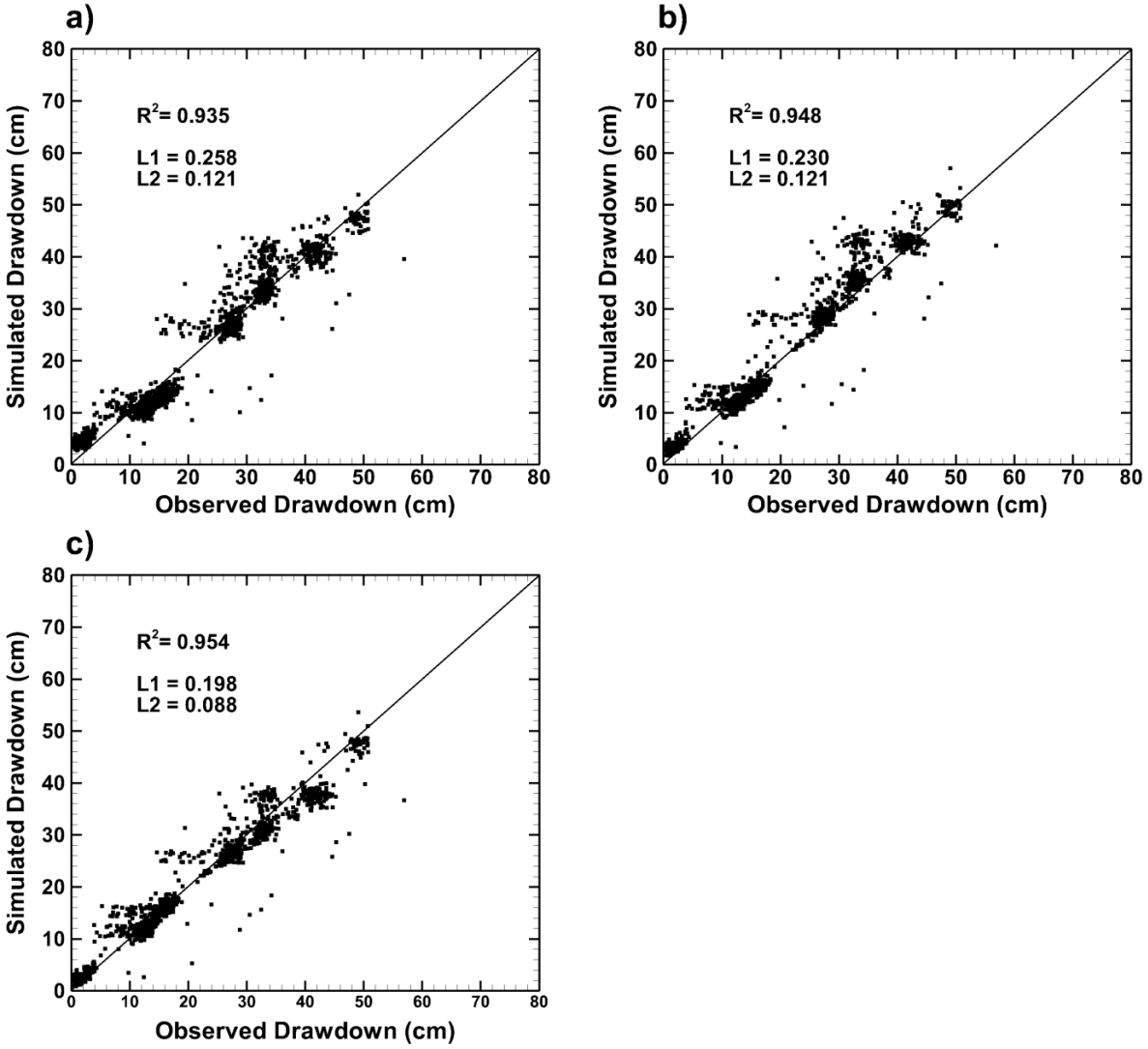


Figure 3.10: Validation scatterplots of observed drawdown and simulated drawdown for: a) homogeneous model, b) HT analysis Case 2, and c) HT analysis Case 1. The scatterplots are shown for data at seven time points (1, 6, 10, 60, 100, 150, 200 min) of all four validation pumping tests (port 13, 35, 37 and 39). The solid line is a 1:1 line indicating a perfect match.

The drawdown curves obtained through the forward simulation treating the medium to be homogeneous are plotted together with predictions based on the results from HT in Figure 3.8 for port 13 [see Supplementary Information section for ports 35 (Figure A1), 37 (Figure A2), and 39 (Figure A3)]. The black solid lines are observed drawdown, while the red solid and blue dashed lines are simulated drawdown based on HT results and homogeneous assumption, respectively. The locations of all ports on Figure 3.8 are kept

the same as the layout on the sandbox shown in Figure 3.1.

According to the simulated drawdown curves for both pressure transducer and tensiometer ports in Figure 3.8, the prediction based on the estimates of HT can accurately capture the drawdown curves for all times at most ports. In contrast, the prediction based on homogeneous assumption fails to capture the s-shaped trend typically observed during pumping tests in unconfined aquifers. Specifically, during the early time (< 3 min) of the pumping test, the prediction based on homogeneous assumption tends to yield larger drawdowns compared to the observed drawdown in almost all the observation ports for all four pumping tests. During the intermediate time (i.e., after 3 minutes), the drawdown predictions based on homogeneous values is smaller than observed values. These results suggest that the homogeneous model fails to capture the drawdown curves at early to intermediate times. However, drawdowns from the homogeneous case match the late time data quite well after the drawdowns have propagated throughout the saturated and unsaturated zones. Therefore, while the homogeneous analysis may provide parameter estimates that could be useful for late-time drawdown predictions, the accurate prediction of drawdowns for all times will require a HT analysis that considers heterogeneity in all parameters. Overall, these results indicate that HT with SLE algorithm is a promising way to characterize the unconfined aquifer such that predicted drawdown response in unconfined aquifer is more representative, compared to those based on homogeneous conceptual model.

Further comparisons of observed and simulated drawdowns at the same time points (1, 6, 10, 60, 100, 150, 200 min) selected from all four validation tests are shown in Figure 3.10a - 3.10c. Meanwhile, the R^2, mean absolute errors (L_1 norm), and mean square errors (L_2 norm) are used to evaluate the simulated and observed drawdown values obtained from different cases: (a) homogeneous model, estimating effective parameters; (b) HT analysis

case 2, only estimating heterogeneous saturated zone parameters, K and S_s; (c) HT analysis case 1, estimating also heterogeneous unsaturated zone parameters in addition to K and S_s. The scatterplots and quantitative statistics show that drawdown predictions for all three cases are unbiased, and progressively improve as more heterogeneity is considered during joint inversion of sequential pumping tests.

3.2.3 Overall Assessment of HT in Unconfined Aquifers

Performing HT to estimate the unconfined aquifer heterogeneity could be a difficult scientific problem when the unsaturated zone is also considered, simply because additional hydraulic parameters are estimated in the unsaturated zone to explicitly capture unsaturated flow, which is highly nonlinear. Aside the work of *Mao et al.* (2013c), previous HT analyses of unconfined aquifers have relied on various simplifying assumptions (e.g. *Cardiff et al.*, 2009; *Cardiff and Barrash*, 2011; *Cardiff et al.*, 2012). In this study, we have validated the new version of HT algorithm (*Mao et al.*, 2013c) that more realistically considers the unsaturated flow process using data from real pumping tests.

The validation results suggest HT could be used for future unconfined aquifer characterization, especially at sites where unsaturated zone flow is critical. On the other hand, the necessary conditions which enable the variably saturated flow inverse problems being well defined should also be considered. This has been investigated by *Mao et al.* (2013a). They stated that, "*... in order to obtain a unique estimate of hydraulic parameters, along each streamlined of the flow field (1) spatial and temporal head observations must be given; (2) the number of spatial and temporal head observations required should be greater or equal to the number of unknown parameters; (3) the flux boundary condition or the pumping rate of a well must be specified for the homogeneous case and both boundary flux and pumping rate*

are a must for the heterogeneous case; (4) head observations must encompass both saturated and unsaturated conditions, and the functional relationships for unsaturated hydraulic conductivity/pressure head and for the moisture retention should be given, and (5) the residual water content value also need to be specified a priori or water content measurements are needed for the estimation of the saturated water content."

For the study here, we preformed the inverse estimation using pressure head data obtained from an unconfined sandbox aquifer drained to certain depths, instead of taking advantage of fully drainage data which will enable us to get more complete drainage curves for the tensiometer pressure transducers. Apparently, the listed necessary conditions are not fully satisfied as only limited temporal and spatial data are used to capture both the saturated and unsaturated zone heterogeneity of the sandbox aquifer. Therefore, it seems that the selection of different initial values of K, S_s, α, and θ_s may pose some uncertainty to the final estimates from the inverse modelling. In order to investigate this issue, two additional cases are performed through changing only the initial values to $K = 2.75 \times 10^{-2}$ cm/s, $S_s = 4.84 \times 10^{-5}$ /cm, $\alpha = 1.6 \times 10^{-1}$/cm, and $\theta_s = 0.33$ (Case 3), which have 48.65%, 22.84%, 23.08%, 2.94% difference compared to values used in Case 1, respectively; $K = 2.63 \times 10^{-2}$ cm/s, $S_s = 5.98 \times 10^{-5}$ /cm, $\alpha = 2.0 \times 10^{-1}$ /cm, and $\theta_s = 0.34$ (Case 4), which have 42.16%, 51.78%, 53.85%, 0% difference compared to values used in Case 1, respectively. Note that $\theta_r = 0.06$ is used in all inverse cases for consistency. Through visual comparison and scatterplots of estimates (see Figures A5 through A8), we find that the general tomogram patterns are basically the same and the estimates of different parameters are close for all three cases. More inversion cases have been done although the results are not provided here since the patterns are similar. On the other hand, changes in L_2 norms are plotted as a function of the number of iterations (see Figure A9) to show the different

calibration behaviours for all three cases. Different initial guesses affect the calibration processes during the early 10 iterations, but then gradually stabilized. Due to the use of different initial values, the computational time differs among cases to reach the stabilized L_2 values: Case 1 (64 iterations over 48 hours), Case 3 (62 iterations over 60 hours) and Case 4 (53 iterations over 36 hours).

Compared to pumping tests in fully saturated aquifers, there are greater complications in analyzing pumping test data from unconfined aquifers through inverse modeling. During pumping in unconfined aquifers, unsaturated zone flow may only be observed over limited time allowing partial drainage of pores, which is also true for our study. Thus limited range of pressure head data were available for characterize this unsaturated zone, and parameters estimated from HT may thus involve great uncertainty. Based on the cross correlation analysis by *Mao et al.* (2013b), head measurements in the unsaturated zone during pumping in the saturated zone carry information about parameters close to the measurement location only, as opposed to a greater region around the measurement location in the saturated zone. In addition, the unsaturated zone parameters, α and θ_s, are more sensitive to late time pressure head data in the unsaturated zone, indicating that inclusion of more late time data from the unsaturated zone will likely better define the non-linear inverse problem in unconfined aquifers. On the other hand, water content data can be used together with the pressure head measurements for the inverse model and more negative pressure head data can be obtained through the installation of additional tensiometer ports in the sandbox aquifer, which will benefit the estimation of the unconfined aquifer parameters. In addition, water content data collected during the drainage process such as with neutron probes or geophysical tools may also be useful in further capturing the unsaturated flow process and further improving the inverse modeling results. However,

these are beyond the work presented in this study, but will lead to future research topics on application of HT in unconfined aquifers.

On the Importance of Geological Data for Hydraulic Tomography: Sandbox Study

This part of the thesis shows the main results of the study in which the importance of geological data is systematically investigated for HT through sandbox experiments. All the pumping test data used in this study are obtained by *Illman et al.* (2010a). Calibration results of different geological models and geostatistical models as well as their prediction performances are presented and compared in the following sections.

4.1 Experimental setup

A synthetic aquifer was constructed in a sandbox by *Illman et al.* (2010a) through the cyclic deposition of different size sediments under varying water flow and sediment feeding rates. For each layer, an uniform flow rate was chosen and the coarser particles appeared to drop first and progressively finer particles deposited on top, producing small-scale heterogeneities within each deposited layers and larger scale heterogeneities of different layers mimicking an interfingering natural fluvial deposit, as shown in Fig. 4.1a. The grain size distribution available from *Craig* (2005) shows that the sands are well sorted. In particular, grain size data in terms of d_{50}, the particle diameter for which 50% of the weight is finer is provided in Table B1 of the Supplementary Information section.

The sandbox has dimensions of 193.0 cm in length, 82.6 cm in height, and 10.2 cm in

width. A series of 4-inch perforated plate/mesh combination was installed at each end of the sandbox to serve as a porous media/water interface and to provide hydraulic control as well as to avoid potential preferential flow. We constructed forty-eight ports along six columns with eight ports each. Each column of ports represents a vertical well instrumented with monitoring intervals. Then, core samples were extracted at each of the 48 port locations and analyzed with a constant head permeameter to obtain local K values. Each horizontal core is 1.28 cm in diameter and 10.6 cm in length. Visual inspections of the cores revealed that the extracted materials are quite uniform and without obvious layering. Therefore, we do not expect significant anisotropy in K within each layer and treat K to be isotropic in all cases presented in this study. Table B1 lists the sand types and K estimates for each layer in the sandbox aquifer. We provide the geometric mean of the calculated K values, when multiple ports intersect a given layer. It is important to note that a sand type itself does not provide a unique K value. That is, a 20/30 sand deposited at various sections of the aquifer all have different average K values. This variation is likely a result of sediment deposition, compaction during the sediment transport process as described, as well as in situ coring and testing of each sample. This is also why that knowing the geological structure does not necessarily mean that we have the accurate K distribution of this synthetic aquifer.

We then conducted twenty-four cross-hole pumping tests at different ports with constant pumping rates that ranged from 2.50 to 3.17 ml/s. During each test, the left, right and top of the sandbox were connected and kept as constant head boundaries of 77.5 cm by ponding water at the top and fixing the hydraulic heads at two ends (*Illman et al.*, 2010a). Meanwhile, we observed hydraulic head changes at all 48 ports until steady state conditions, which were usually established in less than one minute. We then shut off the pump and

monitored the hydraulic heads to full recovery.

We selected steady state head data from 24 cross-hole pumping tests conducted at ports indicated on Fig. 4.1a and chose two study cases using different numbers of pumping and observation data for the purpose of model calibration and validation. In Case 1, we chose eight pumping tests (ports 2, 5, 14, 17, 26, 29, 44, and 47) along two well columns, while drawdown data from 47 observation ports were selected from each test for model calibration. The pumped port for each pumping test was excluded from the inverse analysis because of a skin effect, which is likely to introduce error in the hydraulic head at the pumped well and could have deleterious impacts on parameter estimation (*Illman et al.*, 2007, 2008).

In Case 2, only ports along the second and fifth well column from the left boundary were kept for HT analysis and we chose four pumping tests (ports 26, 29, 44, and 47) and 16 observation ports for calibration (see Figure 4.5 for clear experiment setup). Through this, we present a study case with limited amount of data, that mimics the presence of only two wells screened at various elevations.

Case 3 was identical to Case 2 in terms of the number of pumping tests and monitoring points, but we used geological models of varying accuracy populated with K estimates from permeameter tests as the prior information for the HT analysis. All cases were then validated with 16 independent pumping tests (ports 8, 11, 13, 15, 16, 18 20, 23, 32, 35, 37, 38, 39, 40, 41, and 42) not used in the calibration effort.

The spacing of the pumping and monitoring ports in Cases 2 and 3 does not follow the suggestions by *Yeh and Liu* (2000) for an optimal observation density. However, this case represents the typical field situation more closely than Case 1.

4.2 Description of groundwater models

4.2.1 Forward model

We conducted all steady-state groundwater flow simulations with a variably saturated flow and transport model MMOC3 (for modified method of characteristics) developed by *Yeh et al.* (1993). Both geological and geostatistical models were constructed for the synthetic aquifer which was discretized into 741 elements and 1,600 nodes, with an average element size of $4.1 \times 4.1 \times 10.2$ cm. A finer mesh was also tested, but the results did not vary significantly (*Illman et al.*, 2012). Therefore, for the purposes of consistency with our previous studies that used this discretization (*Illman et al.*, 2010a; *Berg and Illman*, 2011a), we utilized this coarser grid for all cases that we describe in subsequent sections. In terms of boundary conditions, we set the side and top boundaries as constant head, while the bottom, front, and back boundaries of the sandbox were set as no-flow boundaries, as described in the previous section.

4.2.2 Inverse groundwater modeling with the geological zonation approach

We constructed four geological models (*GEO-GOOD*, *GEO-POOR1*, *GEO-POOR2*, and *GEO-POOR3*) to represent the characterization of aquifer layering of different accuracy, shown as Fig. 4.1b through 4.1e. The main purpose of building these different geological models was to examine the impact of accuracy in stratigraphic data on HT analysis. In a different study, *Illman et al.* (2015) utilized a "perfect" geological model to conduct their investigations. This model is also included as Fig. B1 so that it can be compared to the

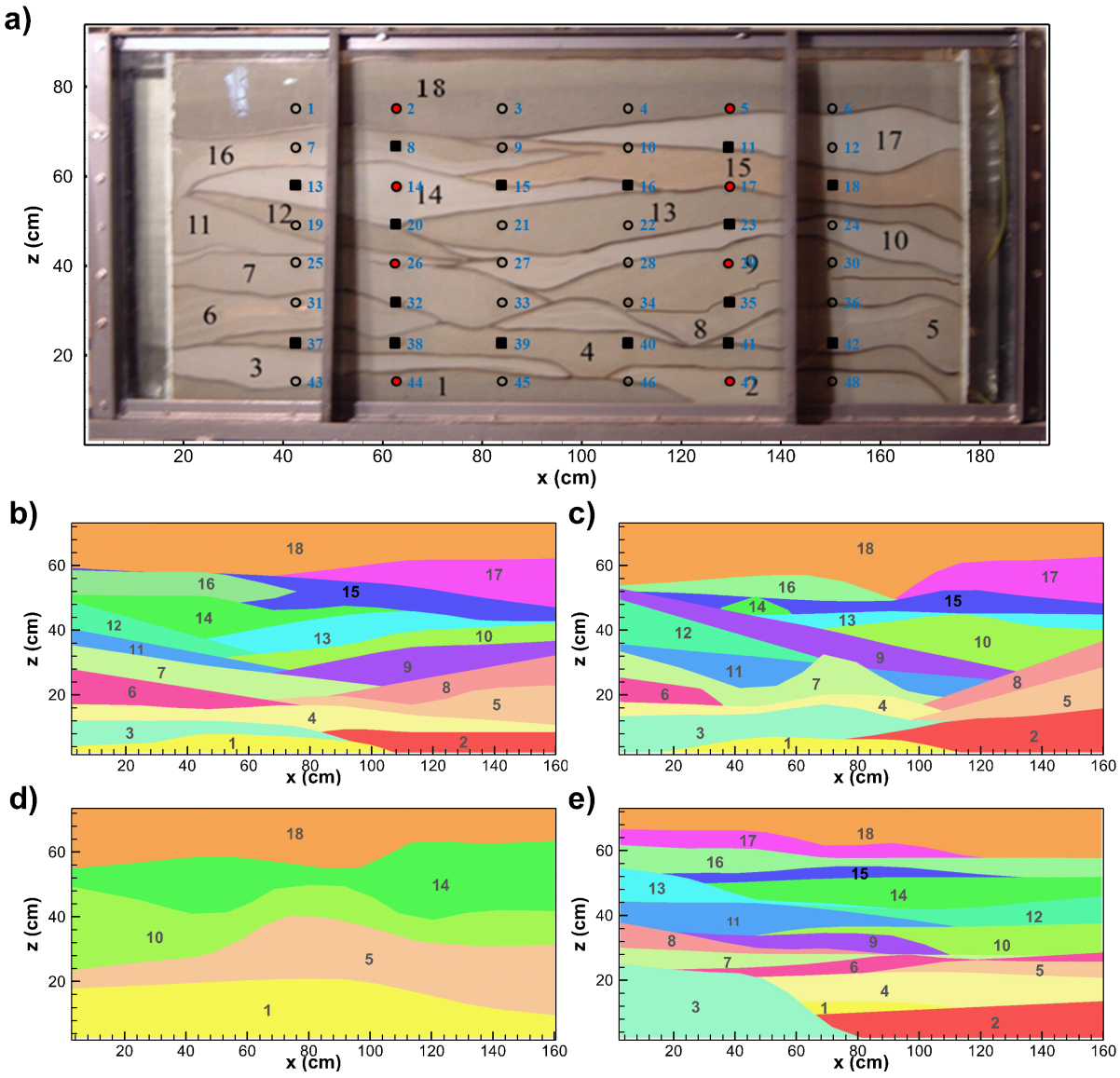


Figure 4.1: (a) Photograph of synthetic heterogeneous aquifer showing the layer (black) and port (blue) numbers (modified after *Illman et al.* (2010a)). Red circles indicate the eight ports (2, 5, 14, 17, 26, 29, 44, 47) pumped for hydraulic tomography, while the 16 black, squares indicate the pumping locations (ports 8, 11, 13, 15, 16, 18, 20, 23, 32, 35, 37, 38, 39, 40, 41, and 42) for the independent cross-hole pumping tests used for validation purposes. (b-e) are geological models built to represent four different descriptions of stratigraphy of different accuracy and layer numbers for: (b) *GEO-GOOD*; (c) *GEO-POOR1*; (d) *GEO-POOR2*; and (e) *GEO-POOR3*.

geological models that we describe below.

The different geological models included borehole information along six columns of wells using Leapfrog Hydro (ARANZ Geo Limited). The underlying algorithm in Leapfrog Hydro is the Fast Radial Basis Function method, which is effectively a way of implementing dual kriging to fill in the gap between different boreholes and create different geological layers. The grid used to create the geological models is identical to the computational grid described earlier.

Among the four geological models, *GEO-GOOD* is constructed to represent the ideal scenario in which the stratigraphy and thicknesses of all 18 layers are assumed to be accurately known along all six columns of wells (Fig. 4.1b). A comparison of the actual stratigraphy on Figs. 4.1a and B1 to Fig. 4.1b reveals that there are considerable differences in terms of layer shapes and how they terminate at the left and right boundaries. The differences in the "perfect" geological model (Fig. B1) of *Illman et al.* (2015) and the *GEO-GOOD* model arises from the fact that in the former, the stratigraphy is mapped directly from Fig. 4.1a, while for the latter, the stratigraphy data along the six wells are interpolated.

For *GEO-POOR1* (Fig. 4.1c), random errors are introduced to the thickness data for layers along boreholes by either arbitrarily increasing or decreasing the thickness values, leading to the obvious deterioration of the interpolated stratigraphy. However, the layering sequences along each borehole are maintained. This kind of error is quite common when collecting and recording core samples during well drilling. Compared to the actual stratigraphy shown in the sandbox (Fig. 4.1a and B1), both the width and thickness of the interpolated stratigraphy for this case varies, especially for the layers in the central part of the sandbox.

GEO-POOR2 represents the scenario with a simplified description of the stratigraphy (Fig. 4.1d). In particular, we merge some layers with similar material types after introducing errors into layer thickness records, generating a simplified geological model with only five zones compared to the actual stratigraphy consisting of 18 layers. Therefore, we consider this case to mimic a scenario in which a geologist has neglected to log fine scale stratigraphic details.

In *GEO-POOR3*, all 18 layers are present (Fig. 4.1e). However, we assume that the geologist provides incorrect stratigraphy information to the groundwater modeller, thus yielding the poorest geological model among the four cases (Fig. 4.1e). Unlike the *GEO-POOR1* model in which the layer sequences down each borehole are correct, the GEO-POOR3 model is constructed based on stratigraphy information with errors even in the locations of some layers. For example, layers 8 and 9 that should be on the right side of the sandbox (Fig. 4.1a), appears on the left side in the GEO-POOR3 model. A comparison of the actual stratigraphy (Fig. 4.1a and B1) and the one in the GEO-POOR3 model shows that only layers 2, 16 and 18 are generally maintained at their original positions.

After the creation of the geological models, we transferred the grid in Leapfrog Hydro to MMOC3. We automatically calibrated each geological model by coupling MMOC3 and the parameter estimation code PEST (*Doherty*, 1994). Similar to the geostatistical inversion that we describe next, all pumping tests were inverted simultaneously amounting to a HT analysis for each of the layer-based geological models. For all layers in each geological model, the K values were treated to be the same and an initial value of K was set as 0.19 cm/s with 1×10^{-4} and 10 cm/s as the minimum and maximum bounds, respectively. Thus, 18 parameters were estimated for the geological model in Fig. 4.1b, 4.1c, and 4.1e, while only five parameters were estimated for the simplified geological model in Fig. 4.1d.

4.2.3 Inverse groundwater modeling with the geostatistical approach

In order to provide direct comparisons to these geological models, we performed geostatistical inversions of all pumping test data during the HT analysis using the Simultaneous Successive Linear Estimator (SimSLE) code (*Xiang et al.*, 2009). One important advantage of SimSLE is that data sets are inverted simultaneously, instead of sequentially, thus providing more constraints for the inverse problem (*Xiang et al.*, 2009).

Previously, the initial value of the unconditional mean was obtained by calculating a geometric mean of K values from the analysis of a pumping test by treating the aquifer to be homogeneous and this value was assigned to the entire simulation domain. Previous studies (e.g. *Liu et al.*, 2002; *Illman et al.*, 2007, 2008, 2010a,b; *Berg and Illman*, 2011a) have shown the use of such a geometric mean value to yield robust HT results. However, research has also shown that the use of a geometric mean value for the entire simulation domain for HT analysis provides less detail to the heterogeneity away from locations where data are available, resulting in smoothed tomogram with K values close to the prior mean (*Illman et al.*, 2015). The main reason for this is that no observation data is available to update the estimates during SimSLE inversion (*Xiang et al.*, 2009).

Here, we present two geostatistical inverse modeling cases. For the inversion starting with a homogeneous field (Cases 1 and 2), the initial values were set as $K = 0.19$ cm/s, $\sigma^2_{lnK} = 3.0$, $\lambda_x = 50$ cm, $\lambda_y = 10.2$ cm and $\lambda_z = 10.0$ cm, which were estimated based on the geostatistical analysis of single-hole test K data, as explained in *Illman et al.* (2010a) and *Berg and Illman* (2011a). In addition, we also used heterogeneous geological models (Fig. 4.1b - 4.1e) as the initial K distributions for the geostatistical inversions (Case 3). In

particular, we assigned initial K values obtained from permeameter tests (Table B1) for each layer of the geological models.

4.3 Model calibration and validation

We performed all calibration runs of geological and geostatistical models on the same PC with a quad-core CPU with 24 GB of Random Access Memory. The eight test calibrations (Case 1) were completed in about half an hour for the geological models, while geostatistical inversions required approximately two hours. For Case 2 using four pumping tests and 16 observation ports, the calibration of geological models required about 15 minutes, while the geostatistical inversion required less than 10 minutes.

To validate the estimated K distributions of calibrated geological and geostatistical inverse models, the 16 pumping tests not used in calibration efforts were simulated for the steady state drawdown at 47 observation ports. Earlier studies by *Illman et al.* (2007) and *Liu et al.* (2007) revealed that the simulation of such independent tests provided the most rigorous quantitative assessment of the estimated parameter distributions.

We assessed the quality of fits between simulated and observed drawdown for both calibration and validation through scatterplots and the best-fit lines and coefficient of determination (R^2) provided on each scatterplot. We also calculated the mean absolute (L_1) and mean square errors (L_2) for each pumping test to further evaluate the fit between observed and simulated drawdowns. Those quantities were computed as:

$$\begin{aligned} L_1 &= \frac{1}{n}\Sigma_{i=1}^{n} \mid \psi_i - \psi_i^* \mid \\ L_2 &= \frac{1}{n}\Sigma_{i=1}^{n}(\psi_i - \psi_i^*)^2 \end{aligned} \tag{4.1}$$

where n is the total number of drawdown data, i indicates the data number, ψ_i and ψ_i^* represent the estimates from the simulated and measured drawdowns, respectively. We calculated the L_1 and L_2 norms for each case by evaluating the observed and simulated steady state drawdowns at each port. We presented both norms as the L_2 norm magnifies large discrepancies and allows one to better assessments of different models.

4.3.1 Case 1: 8 pumping tests and 47 observation ports

Figs. 4.2a through 4.2d show the contour plots of the estimated K tomograms for the four different geological models. In order to facilitate direct comparisons of K tomograms obtained from the inverse models to the K measurements from core samples, we created a K tomogram (Fig. 4.2g) based on the known stratigraphy (Fig. 4.1a and B1) and assigned each layer with K values from permeameter tests.

Examination of Figs. 4.2a through 4.2d reveals that the distribution of high and low K values is quite different among the four geological models due to the use of fixed zones during the calibration process, which clearly shows the compensational effect of estimated parameters for the model structure error (*Refsgaard et al.*, 2012). That is, the calibration process forces the model parameter values to produce simulations as close as possible to observation data under a given structure error in the conceptual model. The overall patterns in terms of the K values for the 18-layer models (*GEO-GOOD* (Fig. 4.2a), *GEO-POOR1* (Fig. 4.2b)) look similar, but because the stratigraphy is different for each model, so are the calibrated K values when they are compared for each layer.

On Figure 4.2a, the best model *GEO-GOOD* that assumes prior knowledge of the stratigraphy, does not preserve the shapes and locations of layers. The layers are not perfectly preserved because of the coarse grid that we use in this study. To maintain the

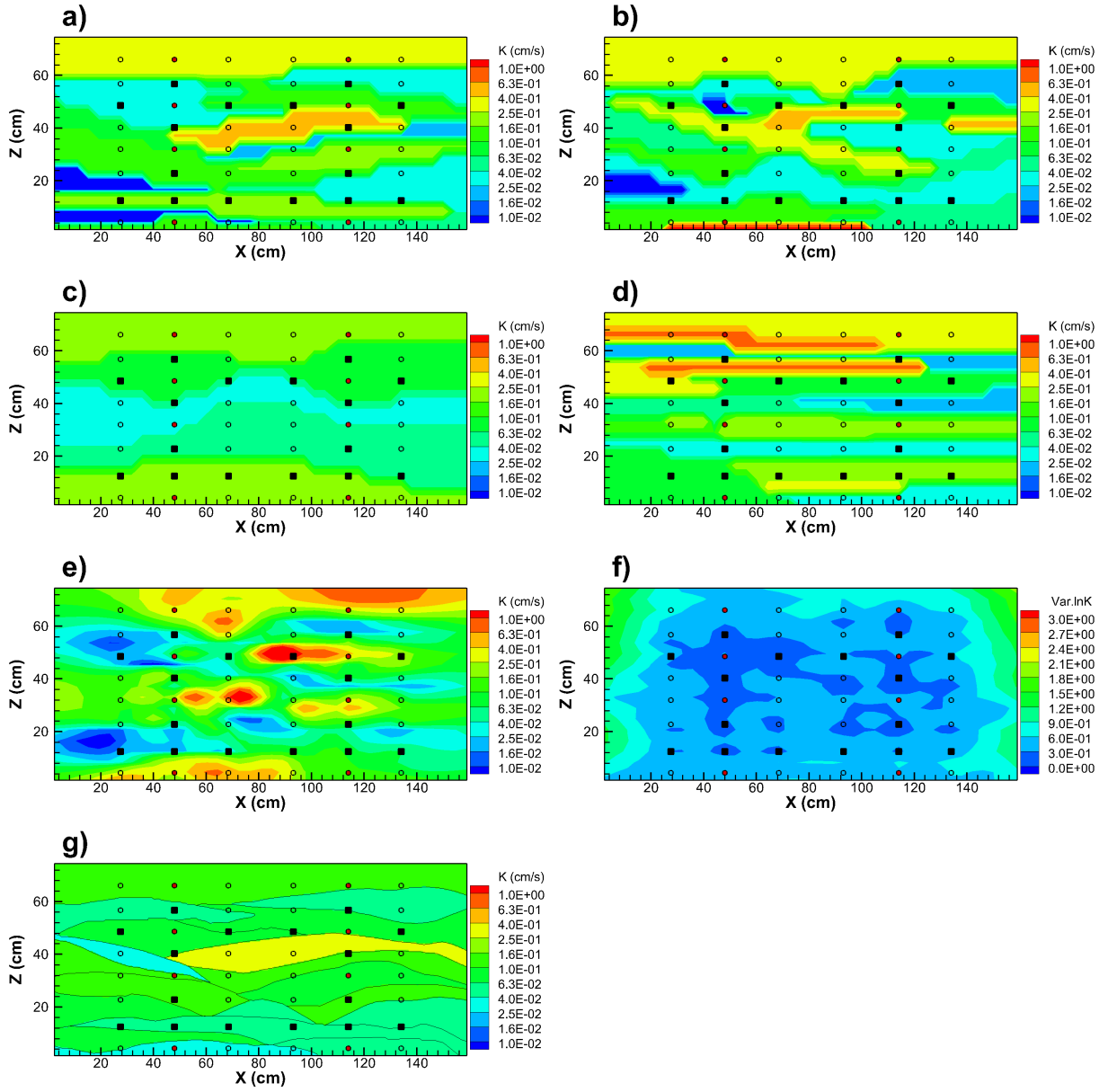


Figure 4.2: Contour plots of different K tomograms (Case 1) using 8 pumping tests and 48 observation ports for calibration: (a) *GEO-GOOD*; (b) *GEO-POOR1*; (c) *GEO-POOR2*; (d) *GEO-POOR3*; (e) SimSLE and (f) residual variances calculated via SimSLE. (g) averaged permeameter K distribution corresponding to the photograph of synthetic heterogeneous aquifer in Fig. 4.1a, with permeameter test K values from Table B1 assigned to each layer.

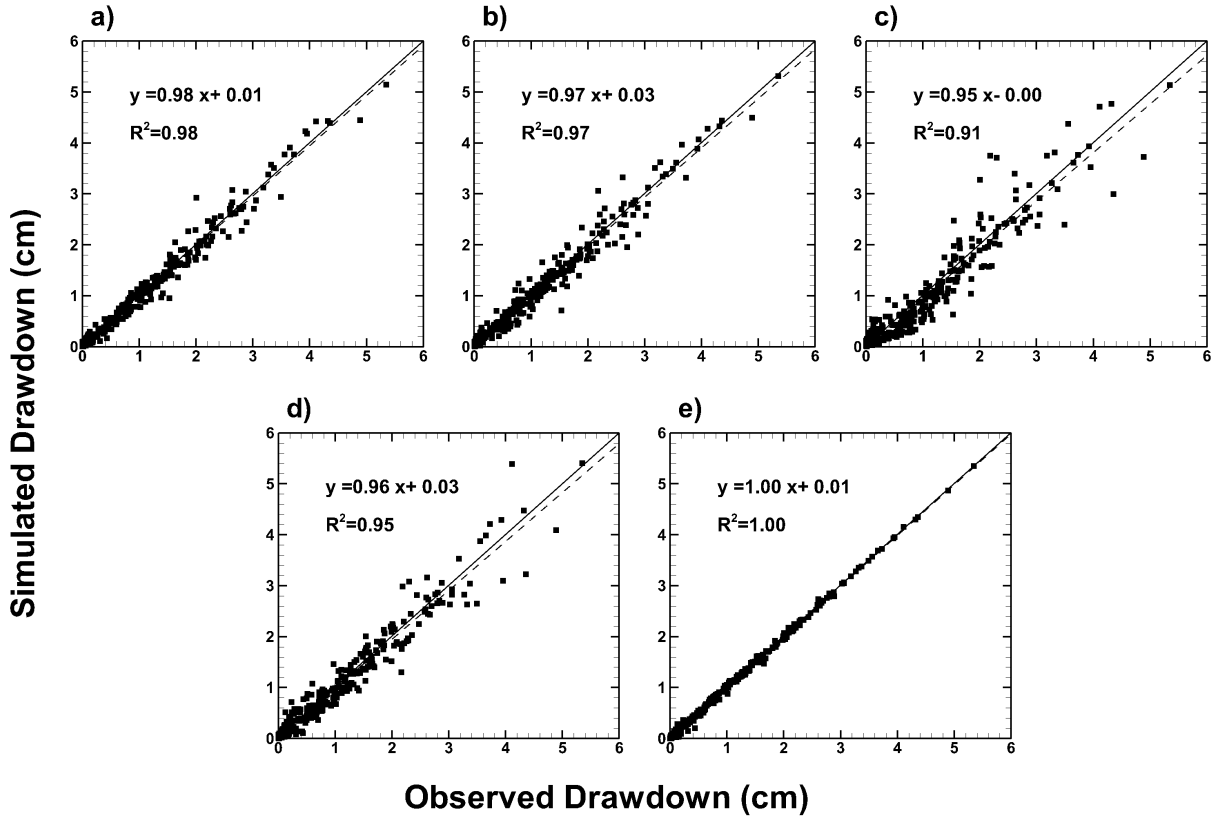


Figure 4.3: Scatterplots of observed versus simulated drawdowns (Case 1) using 8 pumping tests and 48 observation ports for model calibration. (a) - (d) for the four calibrated geological models with different descriptions of stratigraphy: (a) *GEO-GOOD*; (b) *GEO-POOR1*; (c) *GEO-POOR2*; (d) *GEO-POOR3*; and (e) SimSLE.

exact shapes and locations of layers as shown in Figure 4.1b, an even finer grid (In Figure 4.2g, element size is $1 \times 1 \times 10.2$ cm) will be needed for the finite elements instead of the mesh used in the current study (element size is $4.1 \times 4.1 \times 10.2$ cm). To perform fair comparisons, we used consistent grids for both the geological and geostatistical inverse modelling cases.

On the other hand, the simplified geological model *GEO-POOR2* with only five layers (Fig. 4.2c) yields a smoother K distribution than those of the other three geological models with 18 layers. Both high and low K zones visible in *GEO-GOOD* and *GEO-POOR1* are missing in *GEO-POOR2*. Overall, we observe little variation in the calibrated K values for all layers in *GEO-POOR2* and the values of all individual layers are close to the estimated mean K value.

GEO-POOR3 is the worst case in terms of the accuracy in stratigraphy. This case reveals that some high and low K features are evident, but the locations of these layers are quite different from *GEO-GOOD* and *GEO-POOR1* revealing that the poor geological information can have deleterious impacts on inverse modeling. In particular, some parts of the aquifer with low K are shown to have high K and vice versa by comparing Figs. 4.2d and 4.2g. The large differences in the resultant K distributions of the four geological modeling cases are a direct consequence of the trade-off of parameters for fixing the stratigraphy, which are not allowed to adjust during the inverse modeling process (*Refsgaard et al.*, 2012). Figs. 4.2e and 4.2f show the estimated K tomogram and the corresponding ln K variance through the simultaneous inversion of eight pumping test data with SimSLE. Because the layer positions are not fixed and SimSLE estimates the parameters at each grid block, the estimated K distribution shows smooth shapes of high and low K value zones. It is worthy to note that SimSLE yields a K tomogram that captures the major K distribution features

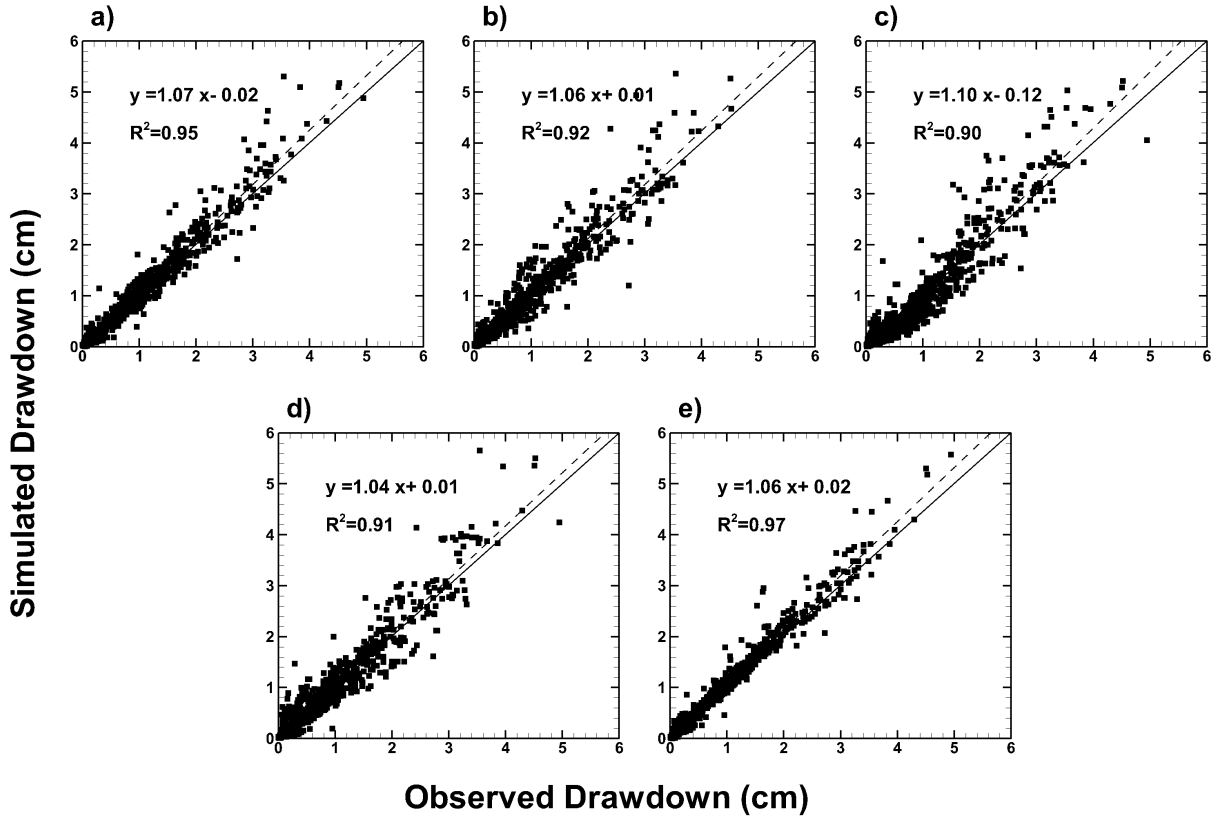


Figure 4.4: Scatterplots of observed versus simulated drawdowns (Case 1) for the 16 pumping tests at the 48 observation ports used for model validation. (a) - (d) for four calibrated geological models with different descriptions of stratigraphy: (a) *GEO-GOOD*; (b) *GEO-POOR1*; (c) *GEO-POOR2*; (d) *GEO-POOR3*; and (e) SimSLE.

shown in Fig. 4.2g. Generally, the distribution pattern of the high and low K zones on Fig. 4.2e corresponds with most of the locations of the various sand bodies (see also Fig. 4.1a and Table B1). However, we note that SimSLE does not capture the precise shapes of the stratigraphic features with this data density. As stated previously, a given sand type does not provide a unique K value from permeameter tests, thus we do not expect the exact correspondence of the estimated K distribution with the stratigraphy. In particular, the K tomogram (Fig. 4.2e) reveals three separate high K zones in the central portion of the aquifer, which is different from the stratigraphy shown on Figs. 4.1a and 4.2g. Although sharp stratigraphic boundaries are not explicitly recovered in the K tomogram obtained from SimSLE, the general correspondence of the K tomogram to the stratigraphy is consistent with previous studies by *Illman et al.* (2007, 2010a,b) and *Xiang et al.* (2009). In particular, *Illman et al.* (2007, 2010b), *Liu et al.* (2007), and *Xiang et al.* (2009) all found that salient features of a nonstationary K field can still be recovered with the sequential (*Yeh and Liu*, 2000; *Zhu and Yeh*, 2005) or simultaneous successive linear approach *Xiang et al.* (2009) that assumes a statistically stationary field. This is because the stationarity assumption becomes less important when a large number of pumping tests and a dense network of observation data are available for geostatistical inverse modeling (*Xiang et al.*, 2009).

We also note that the K values in Fig. 4.2e have a wider range, and hence, a higher degree of K heterogeneity when compared visually to Fig. 4.2g. We later discuss the reliability of the estimated K tomogram computed with SimSLE through model validation.

One significant advantage of using SimSLE for HT analysis is that it yields uncertainty estimates that can aid in assessing the reliability of estimated parameters and to guide future data collection. The ln K variance tomogram (Fig. 4.2f) reveals the uncertainty

in the K tomogram, where the lowest variance values are shown where data is available and near the bottom boundary where drawdown responses are strongest. Uncertainty is highest where data are not available and near the side and top constant head boundaries where drawdowns are weakest.

Figs. 4.3a through 4.3e show the calibration scatterplots of the geological models and those from SimSLE. We provide the corresponding linear model fit, a 45-degree line and R^2 values on each subplot. It is quite evident that SimSLE yields the best calibration result (Fig. 4.3e) with the smallest bias and a least amount of scatter resulting in the highest R^2 value among all subplots. The main reason is that the highly parameterized geostatistical model has a larger degree of freedom to adjust the pressure head data than the zonation-based geological models. Fig. 4.3 also shows that the geological model *GEO-GOOD* (Fig. 4.3a) has a slightly better fit than the poor models, and the simplified model *GEO-POOR2* (Fig. 4.3d) has the worst fit due to lower degree of freedom, indicated by the R^2 values increasing from 0.91 to 0.98 for *GEO-POOR2* and *GEO-GOOD*, respectively. Surprisingly, we find that the poor geological models yield quite good calibration results in current cases where only several K values are adjusted (18 for *GEO-POOR1* and 5 for *GEO-POOR2*) to compensate for structural errors (*Refsgaard et al.*, 2012). However, this does not mean that the calibrated results are reliable and that the K values may be very different from estimates that we obtain locally. Moreover, this does not necessarily mean that the well calibrated poor geological model will perform well in terms of model validation.

We obtained a more quantitative assessment of the calibration results through the calculation of the L_1 and L_2 norms. Fig. B2 in the Supplementary Information section summarizes these results for each pumping test indicated by its port number. We also computed an average value of both norms and ranked the models. The cells of each entry

in the table were color-coded to facilitate an easier comparison of different entries. In particular, we assigned the minimum value in the table a color of dark green, the maximum value a color of dark red, and the 60-percentile value a color of yellow. We utilized a 60-percentile value instead of the median to enhance the contrast in color. These results revealed that SimSLE yields the smallest discrepancy between the simulated and measured drawdowns. In addition, results revealed that SimSLE yields the most consistent results with lowest L_1 and L_2 values across all eight pumping tests. As seen from Fig. B3, *GEO-GOOD* ranked second in terms of L_1 and L_2, *GEO-POOR1* ranked third in terms of L_1 and fourth in terms of L_2, *GEO-POOR3* ranked fourth in terms of L_1 and third in terms of L_2, and GEO-POOR2 ranked fifth in terms of L_1 and L_2. It appears that the simplest model (*GEO-POOR2)* ranked last and the completely wrong model (*GEO-POOR3*) ranked somewhat better throughout all eight pumping tests because there were more adjustable parameters for *GEO-POOR3* to better fit the observed pressure head data.

We also assessed the groundwater models by simulating 16 additional pumping tests not used in the calibration effort and comparing the simulated to observed drawdowns. Fig. 4.4 shows the scatterplot of observed versus simulated drawdowns, while Fig. B3 in the Supplementary Information section shows the corresponding L_1 and L_2 norms. Examination of both Figs. 4.4 and B3 reveals that SimSLE performs the best. We also note that SimSLE provides the most consistent L_1 and L_2 norms for the 16 pumping tests, while these values are less consistent for all geological models. With a high density of monitoring ports and availability of a large number of pumping tests, the estimated K tomogram (Fig. 4.2e) obtained through SimSLE performs markedly better than the best geological model (*GEO-GOOD*) for both calibration and validation, which is consistent with the findings in *Illman et al.* (2015).

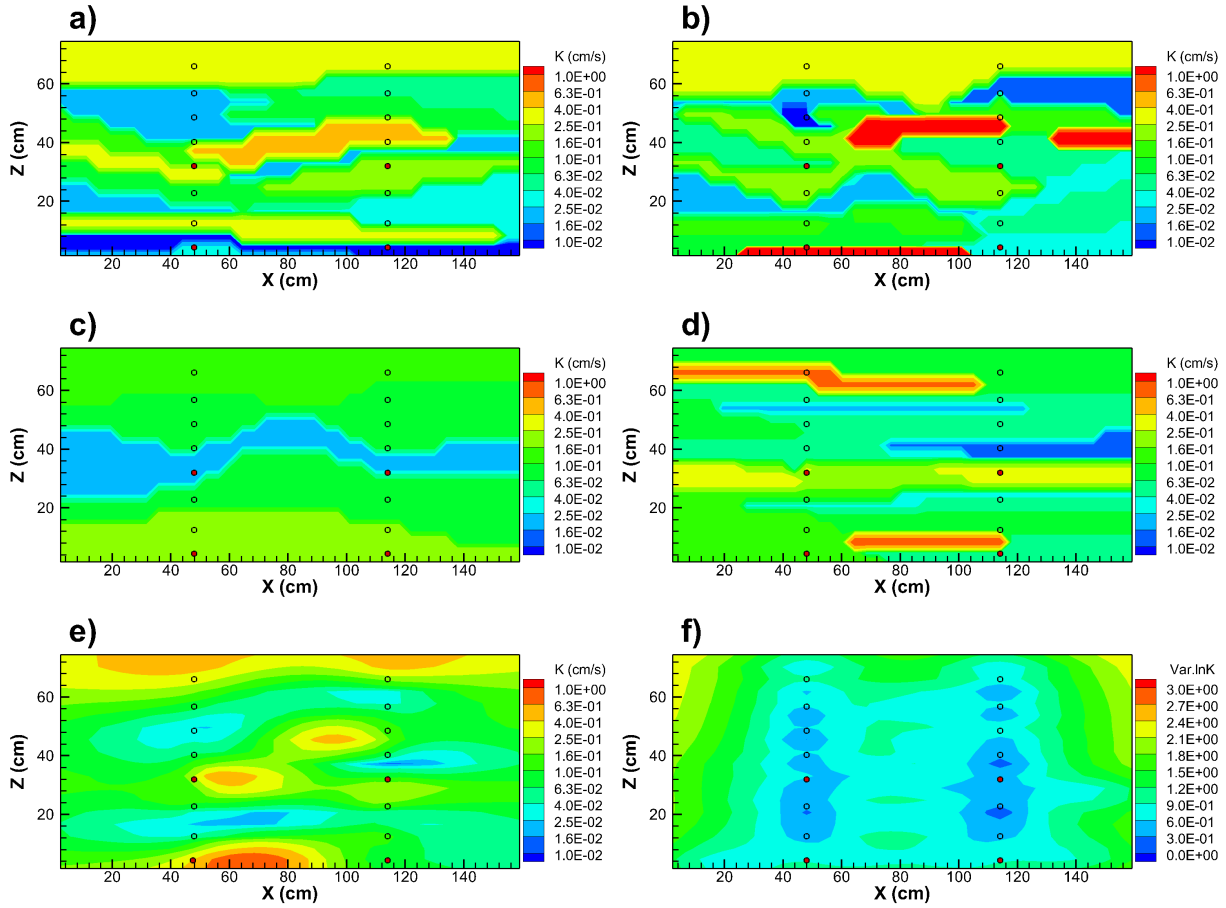


Figure 4.5: Contour plots of different K tomograms for different models (Case 2) using four pumping tests and 16 observation ports for calibration: (a) *GEO-GOOD*; (b) *GEO-POOR1*; (c) *GEO-POOR2*; (d) *GEO-POOR3*; and (e) SimSLE.

4.3.2 Case 2: 4 pumping tests and 15 observation ports

One important aspect related to the potential value of including geological data into HT data interpretation is its calibration and validation performances when a limited number of pumping tests and corresponding monitoring data are available. In order to examine this issue, we reduce the data used for calibration from eight pumping tests to four, and only use observation data from 15 ports placed along two borehole columns instead of the 48 ports for calibration, while we simulated the 16 validation pumping tests and assessed the results at all 48 observation ports. The geological models shown in Figs. 4.1b through 4.1e remain the same and are constructed using core data from six columns of wells as described earlier.

Figs. 4.5a through 4.5d show contour plots of K tomograms through the calibration of each geological model. Comparing to the K tomograms shown in Figs. 4.2a through 4.2d for Case 1 with a larger number of pumping tests and monitoring ports, the patterns of high and low K zones for all geological models are quite different, especially for the geological model GEO-POOR3 (Fig. 4.5d). In contrast, while there are some differences in the K values, the overall pattern for the GEO-GOOD model for Case 1 (Fig. 4.2a) and Case 2 (Fig. 4.5a) are quite similar. This suggests that a smaller number of pumping tests and a limited number of monitoring points do not significantly impact the recovered tomograms when the geological model is accurate. Hence, the collection of accurate geological data is of paramount importance in obtaining good calibration results.

Figs. 4.5e and 4.5f show the estimated K tomogram using SimSLE and its corresponding ln K variance. Compared to Case 1, there is a great loss of detail in the K tomogram for Case 2, while the general K distribution shown in Fig. 4.5e is still similar to that shown in Fig. 4.2e. Additionally, there seems to be a greater loss of heterogeneity in the

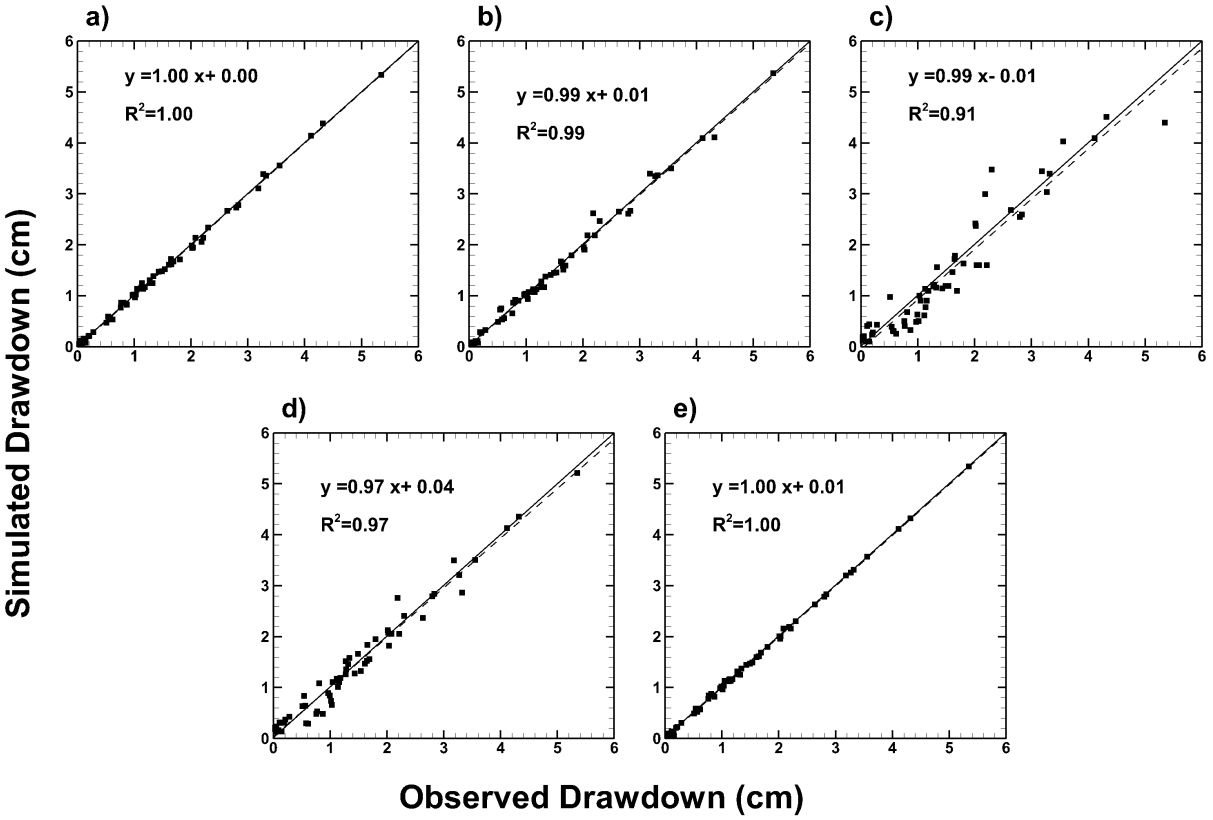


Figure 4.6: Scatterplots of observed versus simulated drawdowns (Case 2) using four pumping tests and 16 observation ports for model calibration. (a) - (d) for four calibrated geological models with different descriptions of stratigraphy and (e) SimSLE.

area outside the two well columns than in between as shown in Fig. 4.5e compared to Fig. 4.2e. The ln K variance (Fig. 4.5f) is higher near the left and right boundaries in comparison to Fig. 4.2f.

The calibration results of Case 2 are also evaluated using scatterplots (Fig. 4.6) as well as L_1 and L_2 norms (Fig. B4 in the Supplementary Information section). Based on these Figures, it is evident that statistics show similar results to Case 1. It is interesting to note that the calibration scatterplots from Case 2 (Fig. 4.6) improve in comparison to Case 1 (Fig. 4.4), due to the fact that considerably less data are used for calibration, while a consistent number of unknowns are estimated for each geological model in Cases 1 and 2. In terms of L_1 and L_2 norms, the ranking for Case 2 remains unchanged compared to Case 1. That is, SimSLE ranked 1, while the rankings of the geological models are

given between the parentheses: *GEO-GOOD* (2), *GEO-POOR1* (3), *GEO-POOR3* (4), and *GEO-POOR2* (5). As explained in Case 1, *GEO-POOR2* ranks last due to the fact that only five layers are calibrated, while the other geological models have 18 layers.

To further examine the model calibration results of the geological models for both Cases 1 and 2, we plotted the K values and their corresponding 95% confidence intervals in Fig. 4.7 with their values provided in the Supplementary Information section (Table B2 for Case 1 and Table B3 for Case 2). The 95% confidence interval is calculated based on the assumption that parameters are normally distributed and on the basis of the linearity assumption used to derive the equations for parameter improvement as implemented in each PEST optimization iteration.

Examination of Fig. 4.7 reveals that estimated K values for some layers have narrow confidence intervals, while others have large intervals. For example, estimated K values from layers 1 through 3 of the *GEO-GOOD* model for Case 2, layer 6 of the *GEO-POOR1* for both Cases 1 and 2, and layers 8 and 13 through 18 for *GEO-POOR3* in Case 2 all exhibit large confidence intervals. The large confidence intervals are mainly due to the stratigraphy being inaccurate and because only few observation ports are available in those layers. Meanwhile, in addition to the varying values between Cases 1 and 2, there are noticeable increases in the confidence intervals of the K estimates for each layer when the amount of data used for calibration is reduced. This is especially true for layer 3 in the *GEO-GOOD* model and layer 6 in *GEO-POOR1* in which no observation ports are available, and for K estimates in layer 7, 8, 13, 16 and 17 of the geological model *GEO-POOR3*.

These results reveal that Case 2 calibration scatterplots for geological models have improved compared to Case 1 in terms of quantitative metrics (R^2, L_1 and L_2), due to the

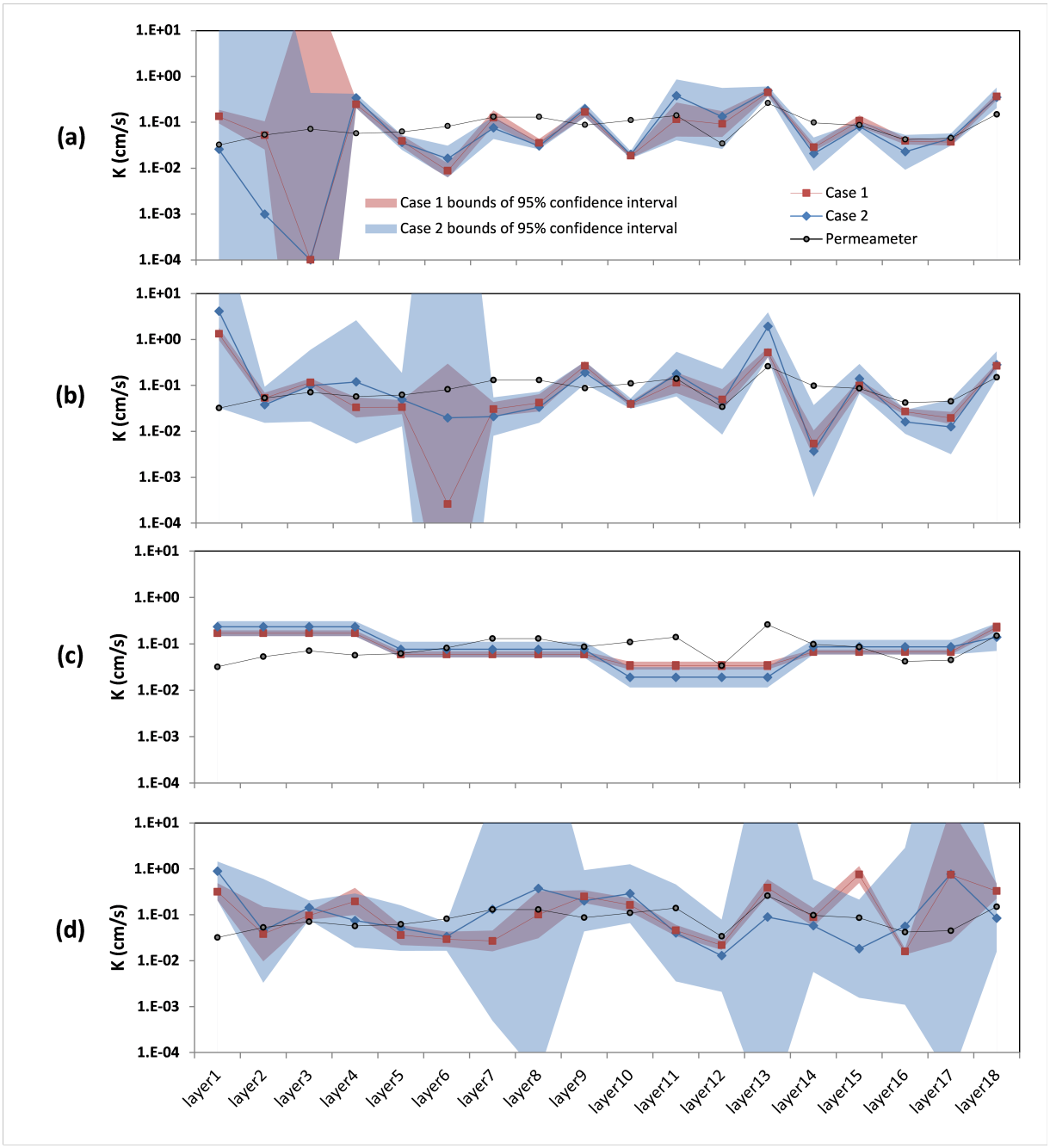


Figure 4.7: Estimated K values and corresponding 95% confidence intervals through PEST calibrations of Case 1 and Case 2 for different geological models: (a) *GEO-GOOD*; (b) *GEO-POOR1*; (c) *GEO-POOR2*; (d) *GEO-POOR3*. In (c), same K values are used for merged layers. Values are provided in Supplementary Tables B2 and B3.

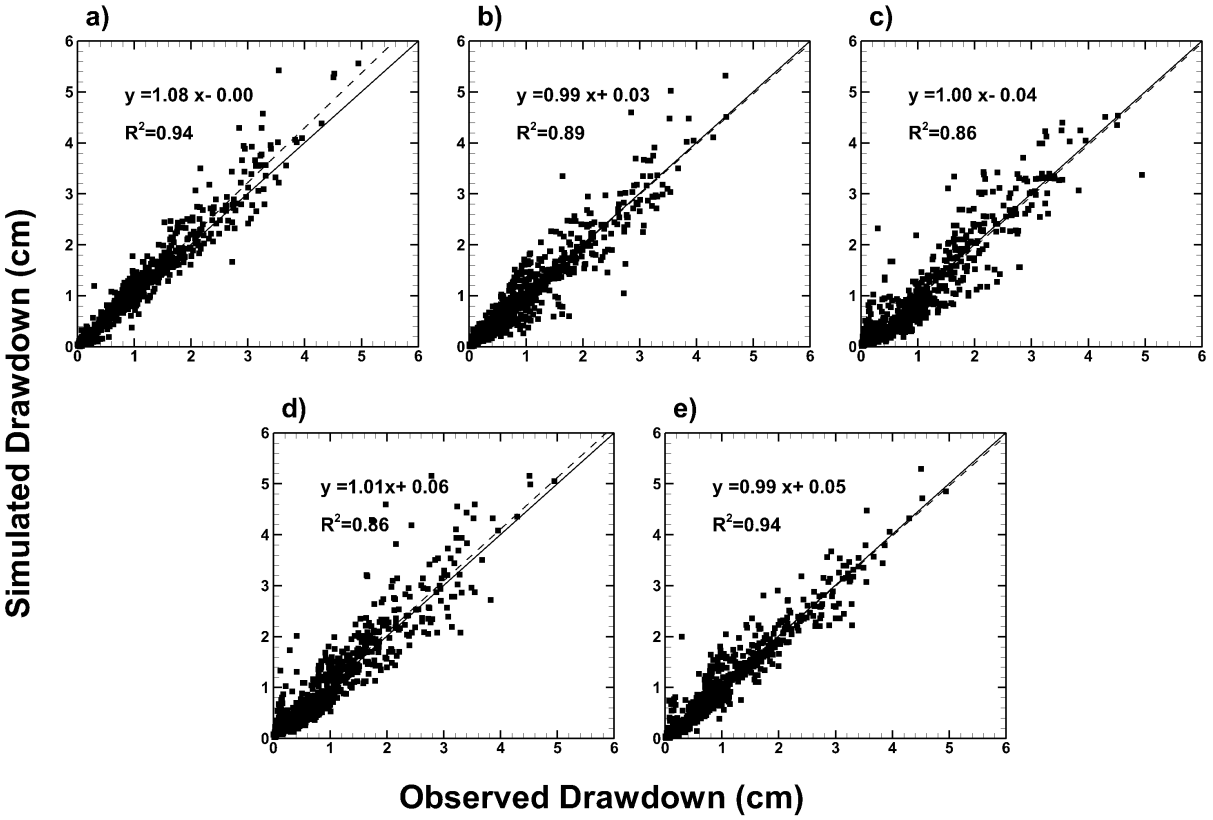


Figure 4.8: Scatterplots of observed versus simulated drawdowns (Case 2) for the 16 pumping tests at the 48 observation ports used for model validation. (a) - (d) for four calibrated geological models with different descriptions of stratigraphy and (e) SimSLE.

use of less data in Case 2 to be fitted by these models. Meanwhile, this also means that in Case 2, less data is used to constrain the inverse model, leading to much wider 95% confidence intervals, suggesting larger uncertainty of estimated values.

Model validation results are discussed next. Fig. 4.8 shows the validation scatterplots, while Fig. B5 in the Supplementary Information section summarizes the L_1 and L_2 norms. Examination of these results reveals that the gaps between SimSLE and the *GEO-GOOD*, *GEO-POOR1* and *GEO-POOR2* models are now much narrower than in Case 1. For example, in terms of the scatterplot (Fig. 4.8), results from SimSLE (Fig. 4.8e) look slightly better with minimal bias in comparison to *GEO-GOOD* (Fig. 4.8a). On the other hand, the L_1 and L_2 norms are very close.

The validation results from the other geological modeling results are about the same

as in Case 1. However, *GEO-POOR3* yielded prediction results that are closer to those of *GEO-POOR2*, which yielded the worst prediction for Case 1 because of the simplified parameterization. We also see from Fig. 4.7d that there are more changes in K estimates for *GEO-POOR3* in Case 2 compared to Case 1, suggesting that the estimates as well as the quality of drawdown predictions in Fig. 4.8 are more likely to be affected by the reduced amount of data for the geological models based on the totally wrong description of stratigraphy. In this regard, proper regularizations may be important and imposed for inverse problems (*Doherty*, 2003; *Carrera et al.*, 2005) to yield more stable estimates. Alternatively, using a simplified model with merged layers of similar aquifer material may be more reliable due to more stable K estimates (Fig. 4.7c). The main reason for this is that, as K values are assumed to be isotropic in our study, the *GEO-POOR2* model merged similar layer material resulting in the estimation of only five variables, while the other three geological models required the estimation of 18 variables. The lower-complexity geological model *GEO-POOR2* will be more likely to be justified than the higher-complexity models with a given amount of observed pressure data (*Schöniger et al.*, 2015), thus leading to more consistent estimates of K values as shown in Fig. 4.7.

Results from Cases 1 and 2 collectively suggest that when few pumping tests and monitoring data are available, accurate geological data become increasingly important for obtaining good results of predicting independent pumping tests not used in calibration effort. Furthermore, the compensational effect of estimated parameters for the model structure errors after calibration is likely to be even more important when the biased model parameter values are used to predict variables beyond the calibration base (e.g., to predict travel time or concentrations), as discussed in *Refsgaard et al.* (2012).

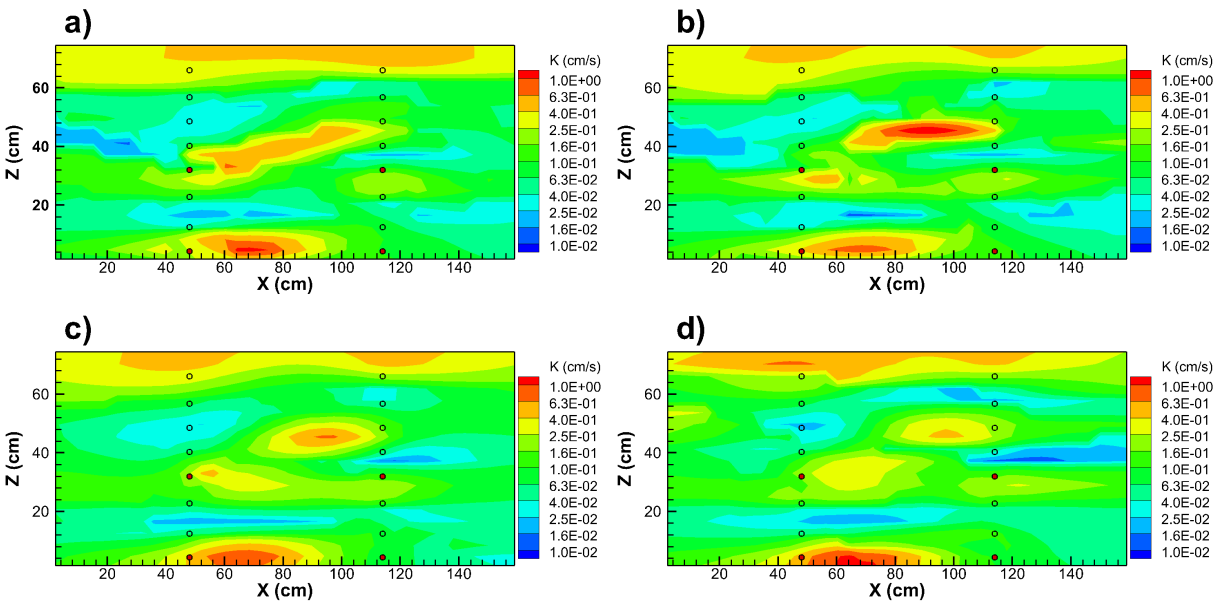


Figure 4.9: Estimated K tomograms obtained via SimSLE inversions (Case 3) using uncalibrated geological models as prior distributions: (a) *GEO-GOOD*; (b) *GEO-POOR1*; (c) *GEO-POOR2*; (d) *GEO-POOR3*.

4.3.3 Case 3: Geostatistical inversions with different prior K distributions

Geostatistical inversions of HT data using SimSLE have proven to be more accurate in drawdown predictions than the perfectly known layer-based geological model in *Illman et al.* (2015). Previously, the geostatistical inversions started with an homogeneous parameter field obtained from available core information or geometric mean values estimated from pumping test data. Using different initial values of K for inversion, *Yeh and Liu* (2000) found minor impacts on the mean value of the estimated parameter field, and the pattern of heterogeneous K distribution remained nearly the same. However, this was for a case with a large number of pumping tests and monitoring data available for inverse modeling. At actual field sites, such high-density data are rarely, if ever available. Previous studies (*Berg and Illman*, 2011b; *Castagna et al.*, 2011; *Mao et al.*, 2013c; *Illman et al.*, 2015; *Zhao et al.*, 2015) have shown that residual variances were higher away from available drawdown

data.

Due to the limited number of pumping tests and lack of high-density monitoring networks at typical field sites, it would be natural to ask whether other data sets could be used to augment monitoring well data. At many sites (e.g. *Sudicky*, 1986; *Boggs et al.*, 1992; *Clement et al.*, 2006; *Sudicky et al.*, 2010; *Alexander et al.*, 2011), geological data, as well as the seismic or ground penetrating radar data could be quite abundant that could provide structural information of subsurface units. Therefore, in order to examine the potential impact of using geological models constructed from borehole data as prior distributions for geostatistical inversions, we utilized uncalibrated geological models (Figs. 4.1b – 4.1e) populated with permeameter K values of each sand type, as shown in Table B1 as initial values for geostatistical inverse modeling with SimSLE. All other initial inputs (e.g., σ^2_{lnK}, λ) and data utilized for inverse modeling were the same as in Case 2 described earlier. The inverse models were run on the same PC and the runtimes were very close to the geostatistical inverse model run for Case 2 with a homogeneous K value as a prior distribution. In order to check whether Case 3 inverse model runs have converged (*Xiang et al.*, 2009), we plotted the changes in L_2 norms obtained from SimSLE runs as a function of the number of iterations (Fig. B8). These results show that all four Case 3 inversions have reached convergence. Fig. B8 also includes the L_2 norms from the geostatistical inversion run from Case 2, which reveals that the calibration process can be affected by the prior distributions during the early iterations, but then gradually stabilizes after 100 iterations suggesting the convergence of SimSLE (*Xiang et al.*, 2009).

The estimated K tomograms obtained from the simultaneous calibrations of four pumping tests (same data as Case 2) using permeameter K value as prior distributions are plotted in Figs. 4.9a through 4.9d. As shown in Figs. 4.9a through 4.9d, the estimated K

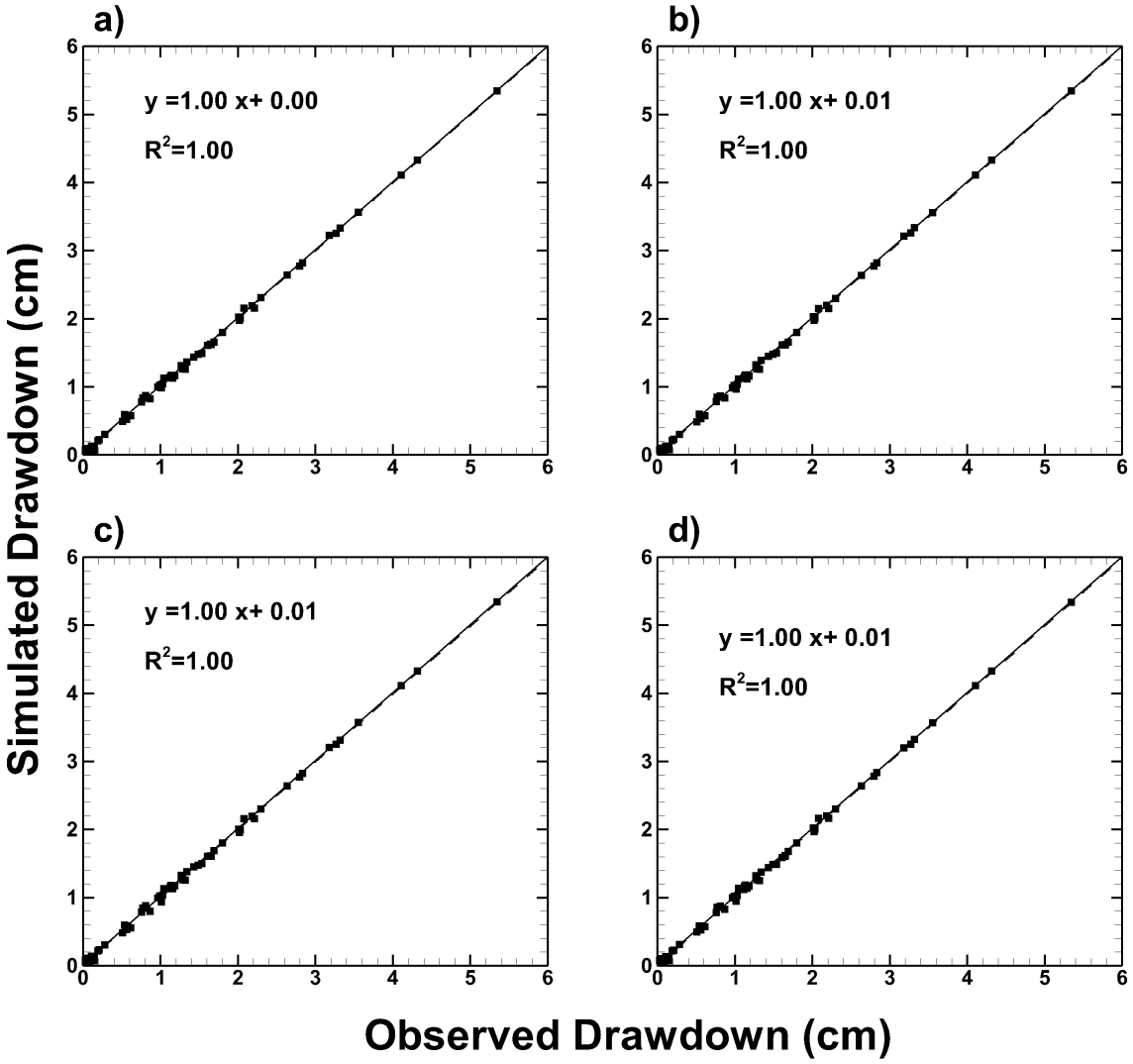


Figure 4.10: Calibration scatterplots of SimSLE inversions (Case 3) using uncalibrated geological models as prior distributions: (a) *GEO-GOOD*; (b) *GEO-POOR1*; (c) *GEO-POOR2*; (d) *GEO-POOR3*.

tomograms are similar in terms of the patterns of high and low K zones, and also similar to the K tomogram obtained for Cases 2 (Fig. 4.5e) obtained with a homogeneous prior distribution, due to using same datasets for calibration. However, the shapes as well as the continuity of these high and low value areas are different from each other. The K tomograms from Case 3 maintain more geological features from the stratigraphy of the deterministic geological models (Figs. 4.2a through 4.2d). These features are obvious for the areas around the boundary and between two well columns, where no monitoring ports are available, especially in Figs. 4.9a, 4.9b, and 4.9d for which the three different 18-layer geological models are used as prior distributions. We also calculate the spatial variances of the recovered K tomograms shown in Figures 4.5e and 4.9a-4.9d. Results show that K tomograms of geostatistical inversion models in Case 3 have slightly larger spatial variances (Figure 4.9a: $\sigma^2_{lnK} = 0.89$; Figure 4.9b: $\sigma^2_{lnK} = 0.88$; Figure 4.9d: $\sigma^2_{lnK} = 0.87$) than the geostatistical inversion model starting with a homogeneous prior distribution in Case 2 (Figure 4.5e: σ^2_{lnK}= 0.75), except for the geostatistical model starting with the 5-layer geological model *GEO-POOR2* (Figure 4.9c: $\sigma^2_{lnK} = 0.74$). The increase in the variance reflects the heterogeneity information provided to the inverse model through the use of available geological data as prior information, compared to the heterogeneity information obtained only from pressure head data.

We additionally compared the observed and simulated drawdowns for model calibration in Figs. 4.10 and Fig. B6, and model validation in Figs. 4.11 and Fig. B7. In Figs. B6 and B7, we also included the geostatistical inversion results from Case 2, marked as SimSLE*, for comparison purposes.

Generally, the SimSLE results using heterogeneous prior K distributions compare favorably among each other and also to the case using an homogeneous prior distribution, in

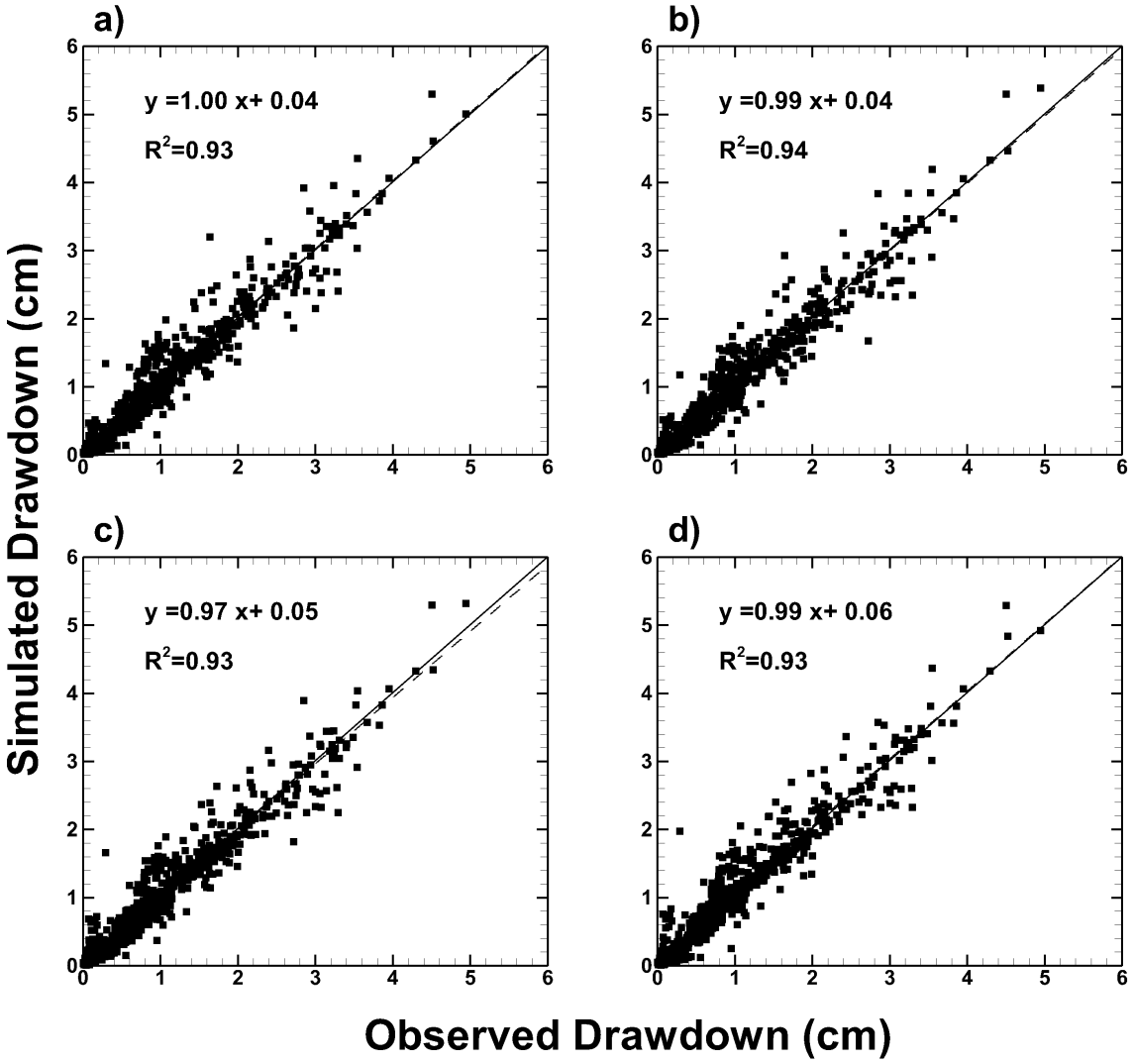


Figure 4.11: Validation scatterplots of SimSLE inversions (Case 3) using uncalibrated geological models as prior distributions: (a) *GEO-GOOD*; (b) *GEO-POOR1*; (c) *GEO-POOR2*; (d) *GEO-POOR3*.

terms of quantitative metrics (R^2, L_1 and L_2) for calibration. When we use the estimated K tomograms to predict drawdowns for individual validation tests, the SimSLE cases with prior information of geology yield minor, but consistent improvements compared to the SimSLE case with homogeneous prior distribution. Most notably, geological structures are visible in these tomograms that are not in the homogeneous initial prior distribution case (Fig. 4.2e). It should be recognized that real geological structures are never known and we used the geological models derived from the same wells used for hydraulic testing in our sandbox analysis. It is of interest to find that the improvements do not seem to impact drawdown predictions of 16 pumping tests significantly, which on one hand, may be due to the diffusive propagation of drawdown perturbation and, on the other hand, to the non-uniqueness of the inverse problem as suggested by *Bohling and Butler* (2010) through the pilot point approach that indistinguishable drawdown responses could be reproduced by hydraulic parameter fields that vary significantly from each other. However, transport predictions may be different as solute transport is affected more significantly by the presence of layer boundaries and fine scale heterogeneities present within layers defined by geologists. However, further investigation of this issue is beyond the scope of this manuscript.

Overall, our sandbox study indicates that improvements to HT analysis are obtained by including geological data as prior distributions for geostatistical inverse models when abundant drawdown data and pumping tests are not available, as clearly shown in Figs. 4.11 and B7.

On the Importance of Geological Data for Three-dimensional Steady-State Hydraulic Tomography Analysis at a Highly Heterogeneous Aquifer-Aquitard System

The laboratory study showed that geological data can be important in HT analysis. However, the laboratory studies are conducted under controlled conditions. Thus, this part of the thesis examines the the importance of geological data for HT at field conditions. In total, 15 pumping tests have been conducted at the NCRS and utilized for this study. Nine pumping tests have been conducted by *Berg and Illman* (2011b) and six tests are conducted by myself for this study.

5.1 Experimental Setup

5.1.1 Site Description

The collection of experimental data took place at the North Campus Research Site (NCRS) located on the University of Waterloo (UW) campus, in Waterloo, Ontario, Canada. Previous Quaternary geology investigations showed that the main deposits in the study area belong to the Waterloo Moraine (*Farvolden et al.*, 1987; *Karrow*, 1979, 1993; *Sebol*, 2000), which consists of multiple glacial till layers resulting from the advance and retreat of ice lobes. During the 1970s, a 30-m long borehole was drilled to investigate the geology beneath

the UW campus. The continuous core samples revealed sequentially deposited tills, from younger to older age, as the Tavistock Till, Maryhill Till and Catfish Creek Till (*Karrow*, 1979; *Sebol*, 2000). In our study area, the surface till is recognized as the Maryhill Till, composed mainly of silty clay accompanied with few stones (*Karrow*, 1979). The younger Tavistock Till only exists rarely and mainly as erosional remnants. Underlying the Maryhill Till is the Catfish Creek Till, which consists of stiff stony silt to sandy silt and is considered to be extremely dense. This till is commonly hard and difficult to drill, thus is referred as the confining unit forming the base of our model. Additional information on the geology of the area could be found in (*Karrow*, 1993).

During our previous work at the NCRS (*Alexander et al.*, 2011; *Berg and Illman*, 2011b, 2013), a nine-well pumping and observation network was developed for HT studies in a 15 m by 15 m area (Figure 5.1a and 5.1b). Four of these wells are continuous multichannel tubing (CMT) observation wells with 10-cm long screens, shown in Figure 5.1b as CMT1, CMT2, CMT3, and CMT4, while the other five are pumping wells (PW1, PW2, PW3, PW4, and PW5). Three wells (PW1, PW3, and PW5) are screened at multiple locations at one-meter intervals. The well screen lengths in these wells are one meter. The other two pumping wells (PW2 and PW4) are nested wells, each including three wells extending to different depths and with one-meter screens.

The site has been investigated through pumping tests and other traditional approaches (i.e., core sampling, permeameter tests, grain size analysis, slug tests) by *Alexander et al.* (2011), *Berg and Illman* (2011b) as well as through this study. Core samples have been analyzed for all four CMT wells and five pumping wells. Soil core sample analyses from the previous and current studies reveal that the main aquifer layers are between seven and 13 meters below ground surface and that this aquifer zone consists of two high K units

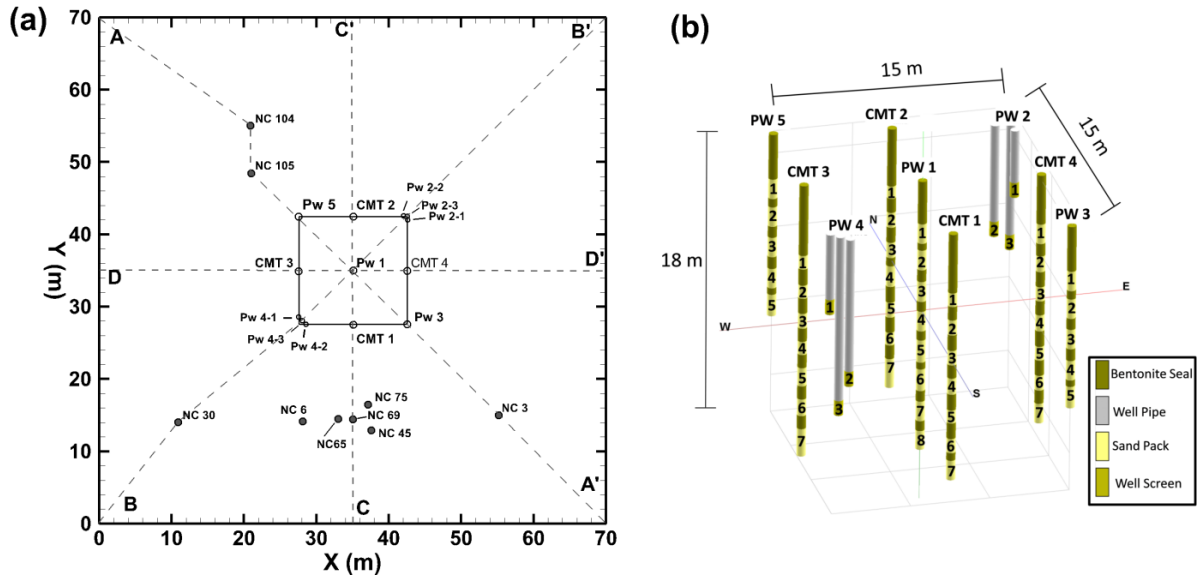


Figure 5.1: (a) Plan view showing well locations at the NCRS situated on the University of Waterloo (UW) campus. Solid filled circles indicate the locations where only geological data are available. Dashed lines indicate four geological cross sections: *A-A'*, *B-B'*, *C-C'* and *D-D'* provided in Figure 5.2. (b) Well screen locations shown for wells clustering in the inner 15 by 15 meter square area where pumping tests are conducted [from (*Berg and Illman*, 2011b)].

separated by a discontinuous low K unit. Drawdown data from pumping tests indicate that the permeable unit behaves as a semi-confined aquifer in our study area. Investigations by *Alexander et al.* (2011) suggested that the low K unit separating the two aquifers is discontinuous and is known to provide hydraulic connections.

5.1.2 Pumping Test Data

Thus far, a total of 15 pumping tests have been conducted at the NCRS and their details are summarized in Table 5.1. Nine pumping tests (PW1-3, PW1-4, PW1-5, PW3-3, PW3-4, PW4-3, PW5-3, PW5-4, and PW5-5) have been conducted by *Berg and Illman* (2011b) of which four tests (PW1-3, PW3-3, PW4-3, and PW5-3) were used for THT *Berg and Illman* (2011b) and SSHT analyses *Berg and Illman* (2013). During the fall seasons of 2013 and 2014, additional pumping and injection tests were conducted at wells PW1-1, PW1-6, PW1-7, PW2-3, PW3-1, and PW5-1, to obtain a more complete hydraulic response data

set for the site.

For this study, we utilized data from 14 pumping/injection tests to perform the SSHT study at the NCRS. During the selection of data for model calibration and validation, only the late time pressure heads from ports indicating steady or quasi-steady state are selected. A quasi-steady state condition is determined from the observed time-drawdown curve in which no significant change in drawdown with time is observed. These quasi-steady state data are mainly obtained from pumping tests that have short pumping/injection durations, such as PW1-1 and PW3-1 as shown in Table 5.1.

Among the selected 14 tests, seven pumping tests (PW1-1, PW1-4, PW1-6, PW2-3, PW3-3, PW4-3, and PW5-3) are used for calibration, while the other seven pumping tests (PW1-3, PW1-5, PW1-7, PW3-1, PW3-4, and PW5-5) are selected for model validation purposes. Pumping tests used for calibration are widely spread over the top and bottom pumping ports, as well as four corner wells in the central 15 m by 15 m pumping and observation area to provide more spatially different flow information for the SSHT analysis. In total, 195 pressure head data are selected for model calibration and 176 head data are used for model validation.

5.1.3 Construction of Geological Models

In order to investigate the value of geological model for SSHT data interpretation, borehole logs of 18 pumping and observation wells are summarized from previous work (*Sebol*, 2000; *Alexander et al.*, 2011) and complied for the construction of the geological model. Figure 5.1a shows the distributions of wells from which geological information are obtained. In total, we used borehole logs from 18 pumping and observation wells completed to different depths, ranging from six meters to 18 meters below ground surface. Based on the soil

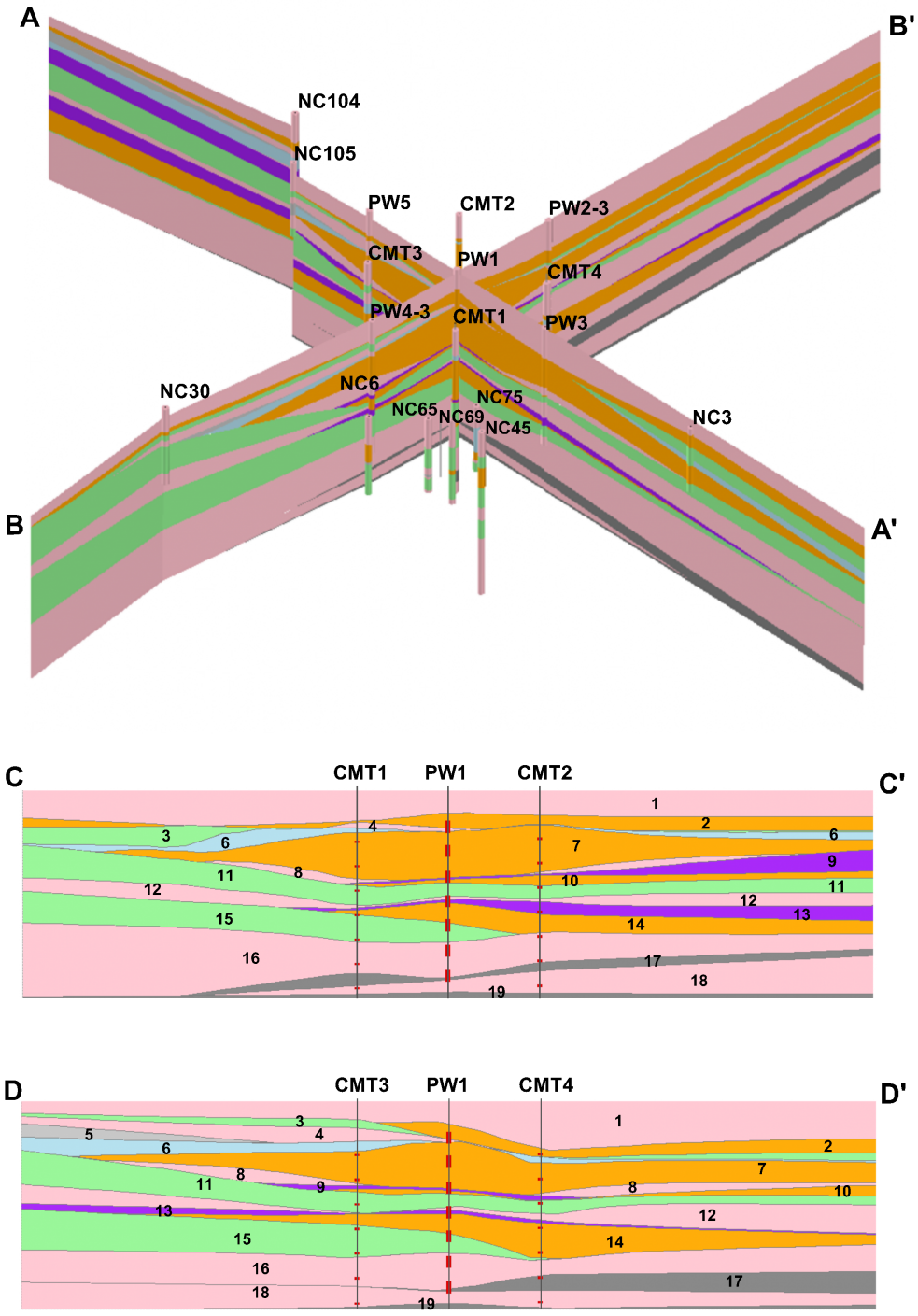


Figure 5.2: Cross sections of the geological model: *A-A'*, *B-B'*, *C-C'* and *D-D'*, at the NCRS. Numbers in cross section *C-C'* and *D-D'* indicate the 19 layers of different materials: Clay (1, 4, 8, 12, 16, 18); Silt and Clay (17, 19); Silt (2, 7, 10, 14); Sandy Silt (6, 9, 13); Sand and Silt (5); Sand (3, 11); Sand and Gravel (15). Screened locations are shown on wells depicted in cross sections *C-C'* and *D-D'*.

Table 5.1: Summary of Pumping/Injection Tests Performed at the North Campus Research Site (NCRS) on the University of Waterloo (UW) campus.

Well Location	Pumping/Injection Rate (L/min)	Duration (hour)	Type
PW1-1	1.89	4.5	Injection
PW1-3	10.5	6.0	Pumping
PW1-4	6.30	8.5	Pumping
PW1-5	4.40	22.5	Pumping
PW1-6	0.95	6.5	Pumping
PW1-7	1.05	26.5	Pumping
PW2-3	1.91	7.0	Pumping
PW3-1	0.94	4.4	Injection
PW3-3	2.10	22.0	Pumping
PW3-4	1.50	22.0	Pumping
PW4-3	30.20	22.5	Pumping
PW5-1	0.85	4.5	Injection
PW5-3	7.80	22.0	Pumping
PW5-4	7.80	8.5	Pumping
PW5-5	8.10	22.0	Pumping

types and corresponding depth information, 19 different layers representing seven different material types are defined along all boreholes.

The layer information between boreholes at different locations are interpolated using the commercial software Leapfrog Hydro (ARANZ Geo Limited), to construct a three-dimensional geological model with dimensions of $70m \times 70m \times 17m$. The underlying algorithm in Leapfrog Hydro is the Fast Radial Basis Function method, which is an effective way of implementing dual kriging that interpolates the stratigraphy between boreholes based on the known geological information from available wells. Four cross-sections (*A-A'*, *B-B'*, *C-C'*, and *D-D'*, in Figure 1a) are extracted along different directions among the central nine wells to illustrate the interpolated geological layers, as shown in Figure 5.2. Moreover, the locations of wells and screens are also presented for cross-sections *C-C'* and *D-D'*.

Based on the layering and soil types, two geological models with different numbers of

layers are prepared for model calibration. One model consists of five layers, while the other more complex model includes all 19 layers. The five-layer geological model is constructed by merging some of the layers with similar material, specifically layers 1 through 10 as layer 1*, layers 12 through 14 as layer 12*, and layers 16 through 19 as layer 16*, which are mainly composed of relatively low K clay and silt. Layers 11 and 15 are treated as two separate zones for the highly permeable nature of sand or sand and gravel. The five-layer model is constructed as a simplified model that only reflects the main high and low permeable zones. On the other hand, the 19-layer geological model is used to take full advantage of the interpolated stratigraphy information. While accurate geological data from 18 boreholes are available for the site, we note that geological information could contain various errors as discussed in a laboratory sandbox study of similar issues by *Zhao et al.* (2016). In particular, *Zhao et al.* (2016) evaluated the impact of utilizing geological models of varying accuracies for HT data interpretations, since the geology was perfectly known. However, in this study, such perfect knowledge of geological information is not available. Therefore, we instead examine the impacts of utilizing two geological models of different resolutions to see whether it improves the HT analyses of pumping test data.

5.2 Description of Inverse Groundwater Models

Three different parameterizations were considered for inverting the HT data in this study: (1) an effective parameter approach by treating the model as homogeneous, (2) a zonation approach based on geological stratigraphy, and (3) a highly parameterized geostatistical approach.

In order to simulate groundwater flow for both forward and inverse modeling, a three-dimensional domain of $70m \times 70m \times 17m$ was discretized into 31,713 computational elements

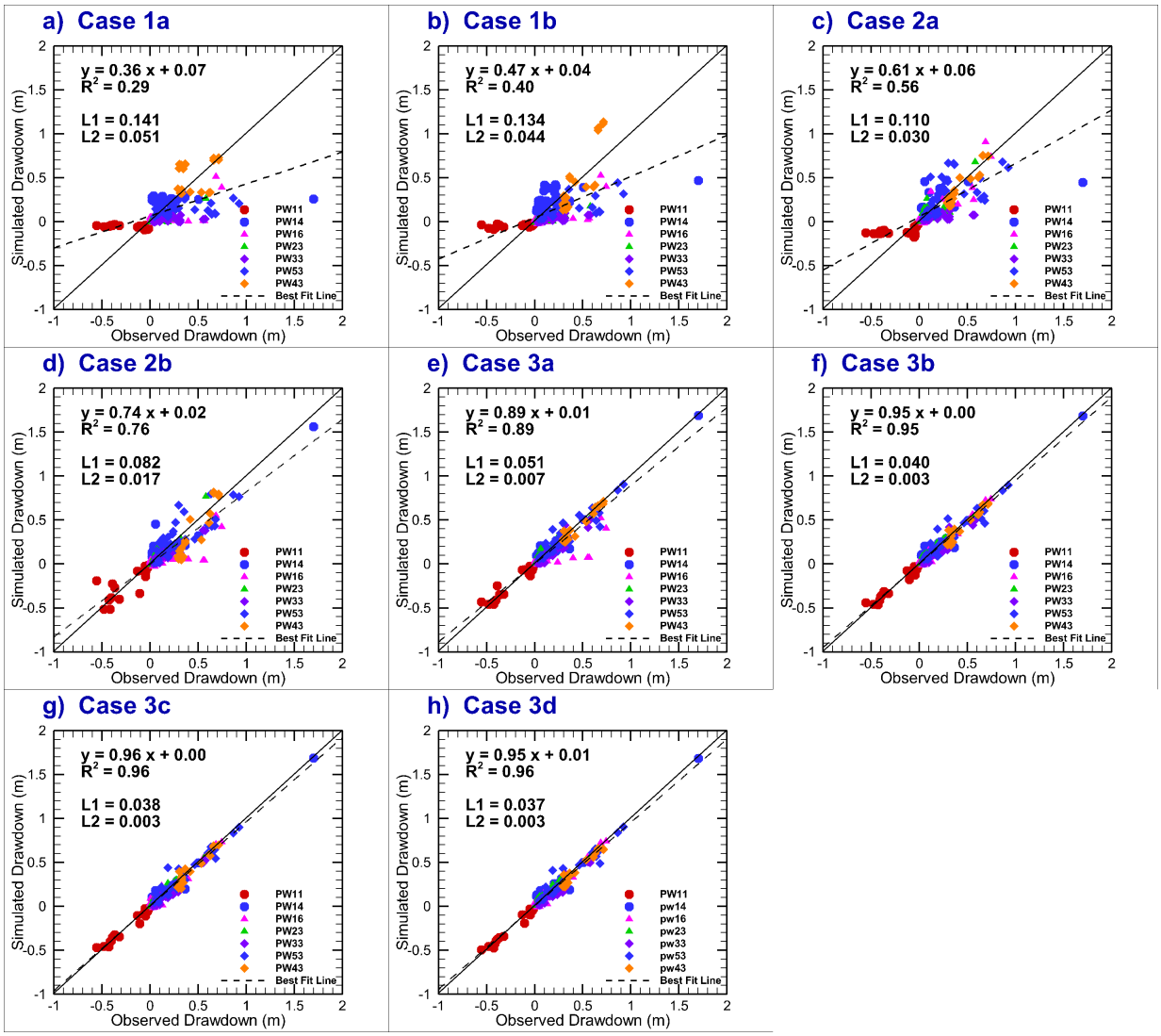


Figure 5.3: Scatterplots of observed versus simulated drawdowns for model calibrations using seven pumping tests for the: (a) isotropic effective parameter model; (b) anisotropic effective parameter model; (c) geological model with five layers; (d) geological model with 19 layers; (e) SimSLE starting with $K = 8.0 \times 10^{-6}$ m/s as prior mean; (f) SimSLE using the calibrated five-layer geological model as prior distribution; (g) SimSLE using the calibrated 19-layer geological model as prior distribution; and (h) SimSLE using the uncalibrated 19-layer geological model assigned with permeameter K values as prior distribution. The solid line is a 1:1 line indicating a perfect match. The dash line is the best fit line. The linear fit results are also included on each plot.

with 34,816 nodes. This grid is similar to the one used previously by *Berg and Illman* (2011b, 2013, 2015) in terms of the general layout, but has a slightly larger simulation domain to include more wells with known borehole data. The elements are gradually refined from the boundary areas to the vicinity of central well locations, decreasing from grid block sizes of $5m \times 5m \times 0.5m$ to $0.5m \times 0.5m \times 0.5m$. The computational grid is provided as Figure C1 in the Supplementary Information section.

All steady-state groundwater flow simulations are conducted using the finite element code MMOC3 (*Yeh et al.*, 1993), which simulates groundwater flow and solute transport under variably saturated conditions. For boundary conditions, the top and bottom faces are defined as no-flow boundaries, while the other four faces are kept as constant head boundaries, as in the previous studies by *Berg and Illman* (2011b, 2013).

5.2.1 Case 1: Effective Parameter Approach

We considered two cases (Case 1a and Case 1b) in the effective parameter approach. Case 1a treats the aquifer to be isotropic where we estimate only K_{eff} and Case 1b treats the entire simulation domain to be anisotropic, for which we estimate the effective K_x, K_y and K_z. Previously, *Berg and Illman* (2015) compared the performance of this approach to the geological modelling approach. However, data from only individual pumping tests were used for calibration. Here, we simultaneously calibrated a total number of 195 pressure heads selected from seven pumping tests. Parameter estimations were performed by coupling the forward simulation code MMOC3 (*Yeh et al.*, 1993) and the model-independent parameter estimation code PEST (*Doherty*, 2005), and pressure head data selected from seven pumping tests were simultaneously calibrated. An initial value of 8.0×10^{-6} m/s with a minimum bound of 1.0×10^{-9} m/s and a maximum bound of 1.0×10^{0} m/s were used

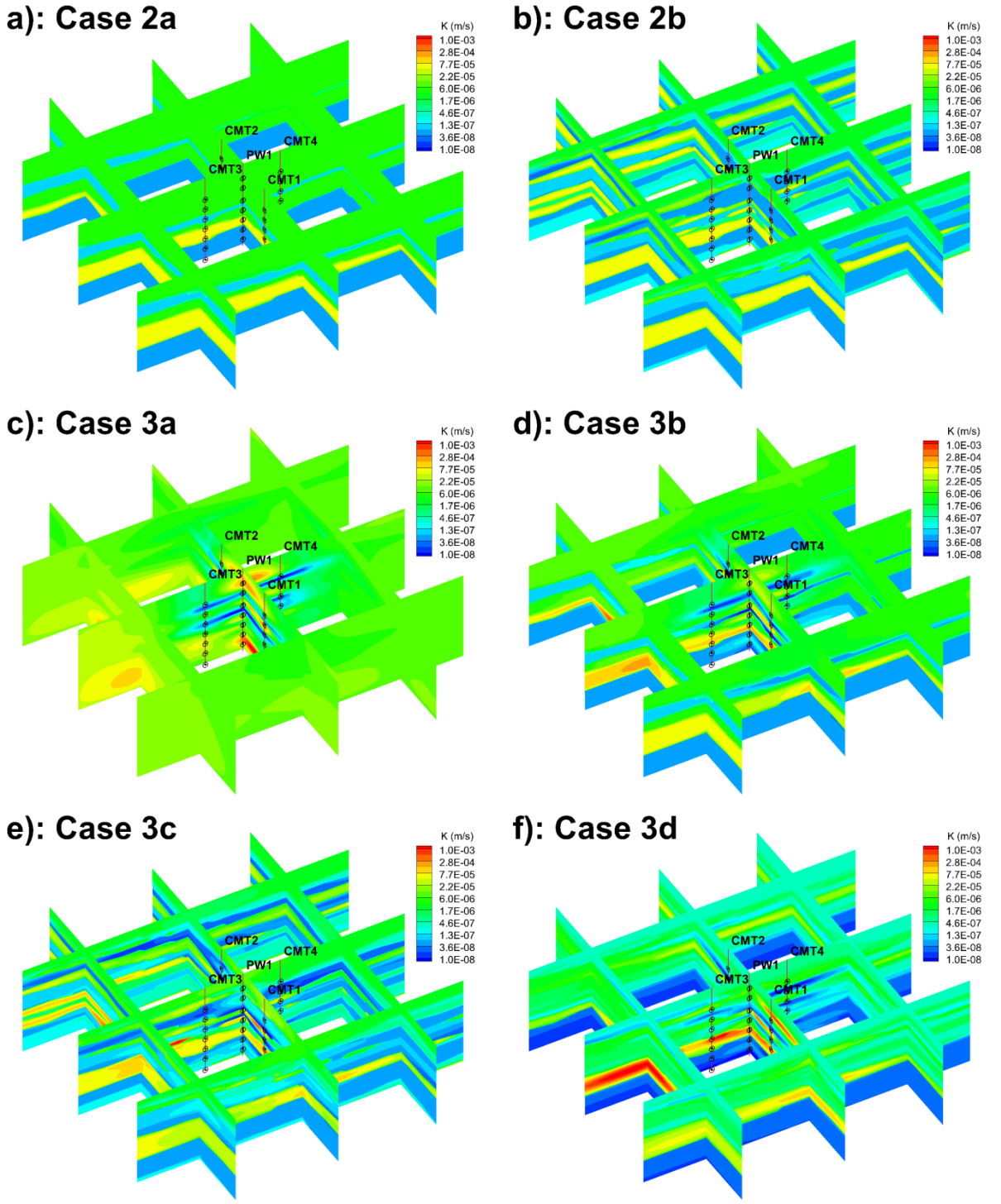


Figure 5.4: Estimated K fields from the inversion of seven pumping tests for: (a) the geological model with five layers; (b) the geological model with 19 layers; (c) SimSLE starting with a uniform $K = 8.0 \times 10^{-6}$ m/s; (d) SimSLE using the calibrated five-layer geological model as prior distribution; e) SimSLE using the calibrated 19-layer geological model as prior distribution; (f) SimSLE using the uncalibrated 19-layer geological model assigned with permeameter test K values for each layer as prior distribution.

for PEST calibration. The initial value that we chose is the geometric mean of individual K estimates obtained by matching the transient drawdown curve at each observation port during pumping at PW1-3 well by *Berg and Illman* (2011b).

5.2.2 Case 2: Geological Zonation Approach

As introduced in the previous section, two geological models were constructed and calibrated: the 5-layer (Case 2a) and the 19-layer geological (Case 2b) models. Both geological models were discretized using the same grid as in the other cases. K values of the elements located in the same layer were treated to be uniform and isotropic. We calibrated both geological models automatically also by coupling MMOC3 (*Yeh et al.*, 1993)) and PEST (*Doherty*, 2005).

The initial K value for calibrating the 5-layer geological model was also set as 8.0×10^{-6} m/s with a minimum bound of 1.0×10^{-9} m/s and a maximum bound of 1.0×10^{0} m/s. While for the 19-layer model, the estimated K values of Case 2a are used as initial values for PEST calibration, in order to speed up the convergence of the PEST run. Results have revealed that the inversion takes approximately 10-hours less when the 5-layer model results are used as the input for the 19-layer model. We also found that this leads to improved correspondence of the estimated K values with available high resolution permeameter K data especially in the lower part of the simulation domain. The main reason for this is the difficulty in obtaining enough drawdown data from low K clay/silt and clay layers located at the bottom of the simulation domain, while pumping from wells located in the top layers (PW1-1, PW1-3, PW3-1, etc.) of the aquifer-aquitard system at the NCRS.

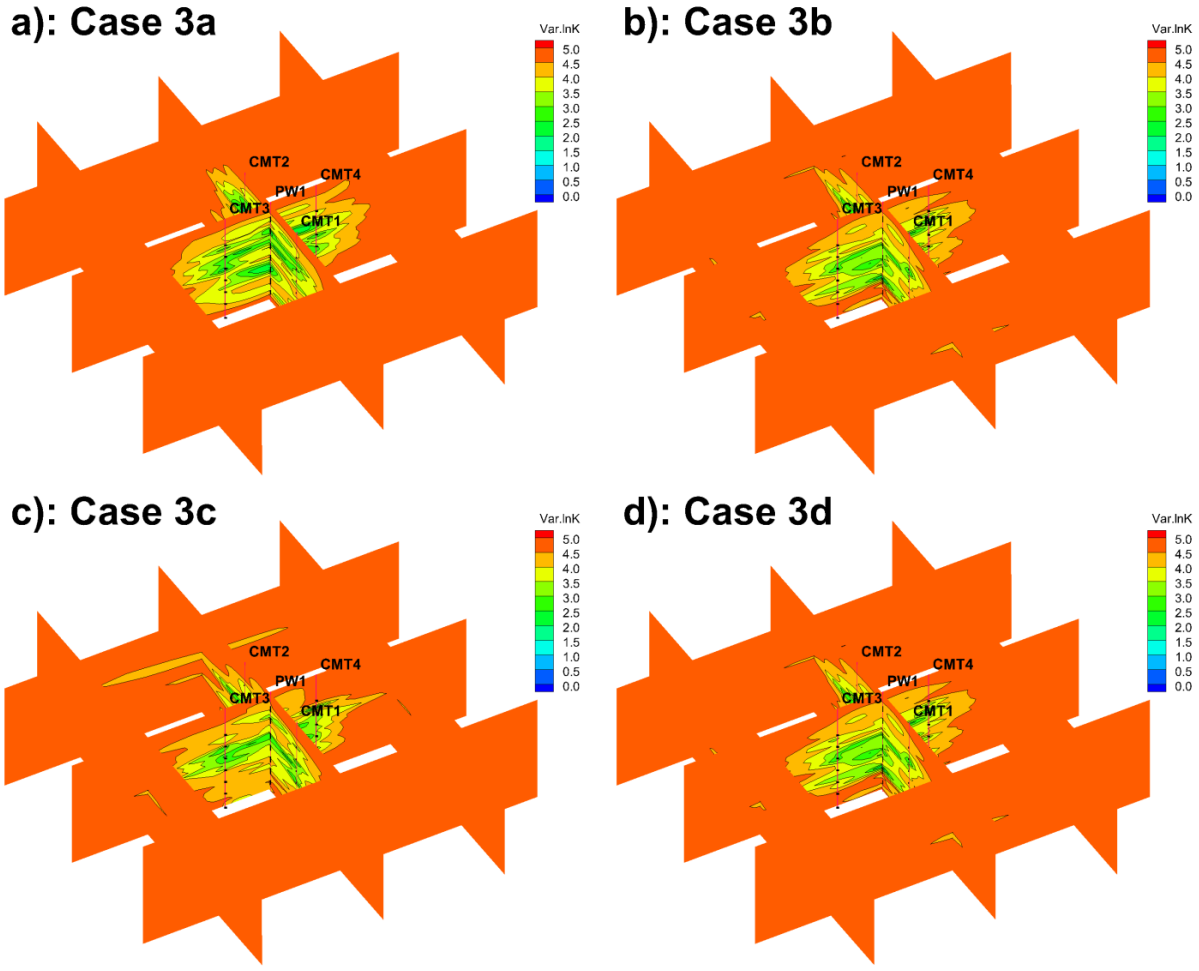


Figure 5.5: Corresponding residual variances of estimated ln K fields from the inversion of seven pumping tests for (a) Case 3a: SimSLE starting with an uniform $K = 8.0 \times 10^{-6}$ m/s; (b) Case 3b: SimSLE using the calibrated 5-layer geological model as prior distribution; (c) Case 3c: SimSLE using the calibrated 19-layer geological model as prior distribution; and (d) Case 3d: SimSLE using the uncalibrated 19-layer geological model assigned with permeameter test K values for each layer as prior distribution.

5.2.3 Case 3: Geostatistical Inversion Approach

The geostatistical inverse modeling of the same pumping test data as in the previous two cases were conducted using the Simultaneous Successive Linear Estimator (SimSLE) code (*Xiang et al.*, 2009), which inverts all the data sets simultaneously, thus providing more constraints to the inverse problem (*Xiang et al.*, 2009) compared to when the data are sequentially included in the inverse code (*Yeh and Liu*, 2000; *Berg and Illman*, 2013). In SimSLE, natural log values of a hydraulic conductivity (i.e., ln K) in the heterogeneous field are treated as a stochastic process, and the stochastic conditional means of these parameters are used for groundwater flow modelling in the aquifer.

The inversion process starts with cokriging using available measurements of hydraulic property and pressure heads to produce the conditional property field, with the assumptions that the unconditional means, spatial covariance functions and structure parameters (correlation scales λ_x, λ_y, λ_z and the variance, σ^2_{lnK}) of hydraulic parameters are known. In this study, the exponential covariance model is adopted for the estimated parameter field. The initial guesses for correlation scales of the K field are set as $\lambda_x = \lambda_y = 4$m, and $\lambda_z = 0.5$m, and the variance is set to be $\sigma^2_{lnK}) = 5.0$, which are the values used by *Berg and Illman* (2011b). The cokriged parameter field is then iteratively updated by SimSLE to minimize the differences between observed and simulated heads.

Four scenarios (Case 3a, 3b, 3c and 3d) are considered for the geostatistical inversion approach to meet our study purposes. In Case 3a, the inversion starts with a uniform mean field of $K = 8.0 \times 10^{-6}$ m/s, which is the same as the initial K value used in the effective parameter and geological zonation approaches. On the other hand, for the other three cases (Cases 3b – 3d), geologic information is used as prior knowledge for the inversion. Specifically, Case 3b uses the estimated K values from Case 2a as the prior mean

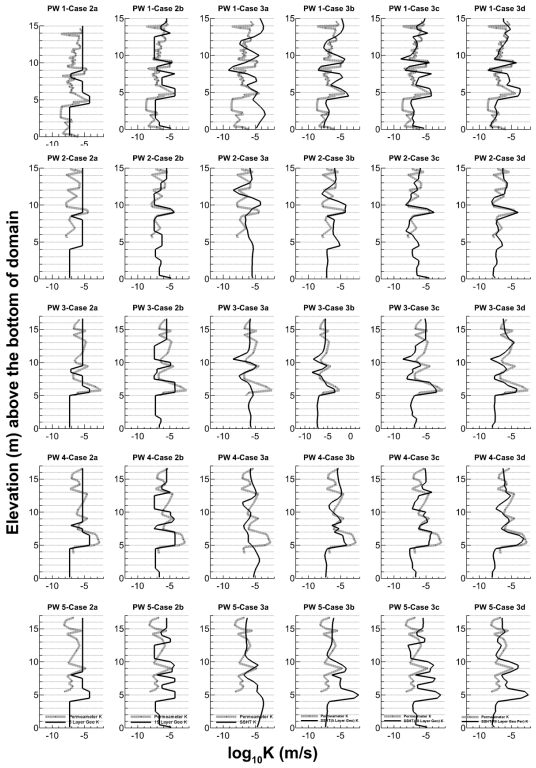
distribution; Case 3c uses the K estimates from Case 2b as the prior mean distribution; and Case 3d uses the 19-layer geological model (Case 2b) populated with permeameter tested K values as the prior mean distribution. In Case 3d, the corresponding K values for each layer are calculated as the geometric mean of soil sample measurements located in the same layer and these values are listed in Table 5.3. In Table 5.3, permeameter test K values of layer 5 and layer 19 are estimated to be the same as the values of layers 6 and 17, respectively, due to the fact that no core samples are available for layers 5 and 19, but having similar soil material with layer 6 and 17.

Through Case 3b, 3c and 3d, we test the impact of using both calibrated geological models and permeameter test K values as prior mean K distributions for the geostatistical inversion approach. Thus, the findings of Case 3b and 3c would be more useful for practitioners than Case 3d, since the collection of high resolution permeameter test results will require considerable efforts.

5.3 Model Calibration

Inverse modeling of steady state data from seven pumping tests were performed on the same PC with a quad-core CPU and 24 GB of RAM for the effective parameter and geological zonation approaches. In terms of computational time, Case 1a took less than 15 minutes and Case 1b took approximately one hour. Similarly, Case 2a took about 4.5 hours to estimate five unknown values, while Case 2b took about 40 hours for 19 unknown K values. The long computational time is a direct result of running both forward simulation and PEST optimization sequentially using a single CPU, which could be greatly reduced if a parallel computing environment is implemented. On the other hand, geostatistical inversions were performed on a PC-cluster using 16 processors with 192 GB of RAM. Due

(a) PW wells



(b) CMT wells

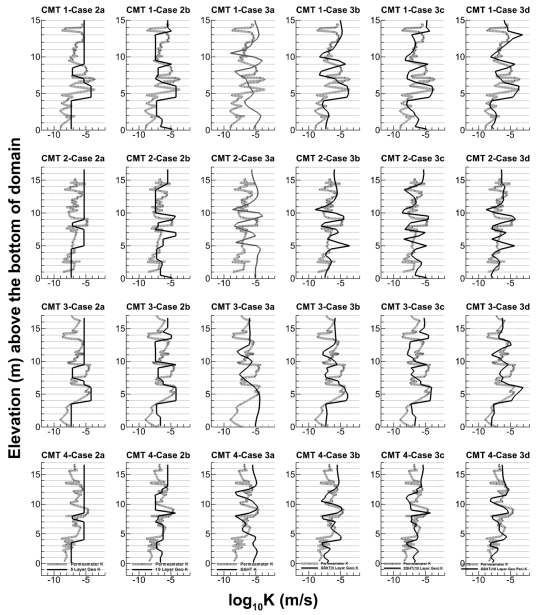


Figure 5.6: Vertical $log_{10}K$ profiles along nine boreholes of (a) PW wells and (b) CMT wells, for different calibration cases. Case 2a: the 5-layer geological model; Case 2b: the 19-layer geological model; Case 3a: SimSLE starting with an uniform $K = 8.0 \times 10^{-6}$ m/s; Case 3b: SimSLE using the calibrated 5-layer geological model as prior distribution; Case 3c: SimSLE using the calibrated 19-layer geological model as prior distribution; and Case 3d: SimSLE calibration case using the uncalibrated 19-layer model assigned with permeameter test K values for each layer as prior distribution.

to the highly parameterized nature of the model, geostatistical inversions took about two hours for one SimSLE iteration, including time for both forward simulation and inverse estimation of hydraulic parameters. In the current version of SimSLE, the inversion part runs in parallel, but the forward model runs on a single CPU. Thus, the geostatistical inversion approach took much longer than the other approaches, but all cases converged within seven days.

To quantitatively assess the performance of model calibration and validation results of all inversion models, the mean absolute error (L_1) and mean square error (L_2). The calibration performances of all models were first evaluated by comparing the linear fit between the simulated and observed pressure heads and overall L_1 and L_2 norms for seven pumping tests. Figure C2 in the Supplementary Information section summarizes the L_1 and L_2 norms of model calibrations as well as the corresponding ranks for all cases. The cells of each entry in the table were color-coded to facilitate an easier comparison of different entries. In particular, we assigned the minimum value in the table a color of green, the maximum value a color of red, and the 60-percentile value a color of yellow. We utilized a 60-percentile value instead of the median to enhance the contrast in color.

5.3.1 Case 1 Results

The simultaneous calibration of the effective parameter model with data from seven pumping tests for the isotropic Case 1a yielded an estimated K_{eff} of 8.4×10^{-6} m/s and a corresponding uncertainty indicated by the 95% confidence interval, which has an upper limit of 9.8×10^{-6}m/s and lower limit of 7.2×10^{-6} m/s. For the anisotropic Case 1b, K_x was estimated as 1.04×10^{-5} m/s with an upper limit of 1.54×10^{-5} m/s and a lower limit of 7.02×10^{-6} m/s, and K_y was estimated as 1.19×10^{-5} m/s with an upper limit

of 1.68×10^{-5} m/s and a lower limit of 8.36×10^{-6} m/s. The effective K values in the horizontal directions x and y are similar. The value of K_z was two-orders of magnitude lower than K_x and K_y, estimated as 6.37×10^{-7} m/s with an upper limit of 1.08×10^{-6} m/s and a lower limit of 3.75×10^{-7} m/s.

When treating the heterogeneous aquifer to be uniform, the estimated parameters are found to be dependent on observation as well as pumping locations (e.g. *Wu et al.*, 2005; *Straface et al.*, 2007; *Wen et al.*, 2010; *Huang et al.*, 2011; *Sun et al.*, 2013; *Berg and Illman*, 2013, 2015). The previous THT study by *Berg and Illman* (2015) found that the effective parameters varied depending on the location of pumping tests when estimating these values for each pumping test at NCRS. Therefore, the estimated effective K values from Case 1a and 1b should be more representative of the test area in an average sense, since the effective K values are estimated by simultaneously considering data from all seven pumping tests.

Examination of the calibration scatterplots (Figure 5.3a) shows that, for Case 1a, the linear fit has a slope of 0.36 and a coefficient of determination (R^2) of 0.29 between simulated versus observed drawdowns. The linear fit from Case 1b improves when compared to Case 1a, yielding a slope of 0.47 and a coefficient of determination (R^2) of 0.40, when treating the aquifer to be anisotropic.

In terms of the L_1 and L_2 norms, the calibration results of the effective parameter model ranked at the bottom two (Figure C2): Case 1a ranked 8^{th} and Case 1b ranked 7^{th}. We also notice that the L_1 and L_2 norms vary significantly from one test to another for both cases. These findings suggest that, although it is possible to obtain relatively good fits for some observation data points, overall, the effective parameter model cannot capture the drawdown responses at the NCRS and the effective parameter cannot be deemed to be

Table 5.2: Estimated K values and corresponding posterior 95% confidence intervals for the 5-layer geological model.

Layer	Estimated K (m/s)	95% Confidence Intervals Lower limit	Upper limit
1^{*a}	5.33×10^{-6}	4.22×10^{-6}	6.74×10^{-6}
11	7.74×10^{-8}	3.94×10^{-8}	1.52×10^{-7}
12^{*b}	5.12×10^{-6}	4.47×10^{-7}	5.87×10^{-5}
15	6.38×10^{-5}	4.63×10^{-5}	8.78×10^{-5}
16^{*c}	4.84×10^{-8}	3.01×10^{-8}	7.80×10^{-8}

[a] Layer 1* is a merged layer of the original Layers 1 through 10;
[b] Layer 12* is a merged layer of the original Layers 12 through 14;
[c] Layer 16* is a merged layer of the original Layers 16 through 19.

representative of the site.

5.3.2 Case 2 Results

For this case, the simultaneous calibration of the 5-layer geological model (Case 2a) completed after 172 model calls. The drawdown scatterplot shows that the overall fit for the 5-layer geological model has a slope of 0.61 and a coefficient of determination (R^2) of 0.56 between simulated and observed drawdowns (Figure 5.3c), which is better than the effective parameter cases (Figure 5.3a and 5.3b). The estimated K values and the corresponding 95% confidence intervals are listed in Table 5.2, while the estimated K distribution is presented in Figure 5.4a. Generally, the calibration of the 5-layer geological model yielded the highest K value for the sand and gravel layer (layer 15) and the lowest K value for the bottom merged layer 16*, consisting of silt and clay layers (layer 16 through 19). K estimates for merged layer 1* and 12* are close to the initial value of 8.0×10^{-6} m/s, which may be the result of using a single layer for multiple soil types. In addition, the upper sand layer (layer 11; Figure 5.2) known to have a high K value, was assigned a value of 7.74×10^{-8} m/s suggesting that the layer is a low K zone, which is inconsistent with known

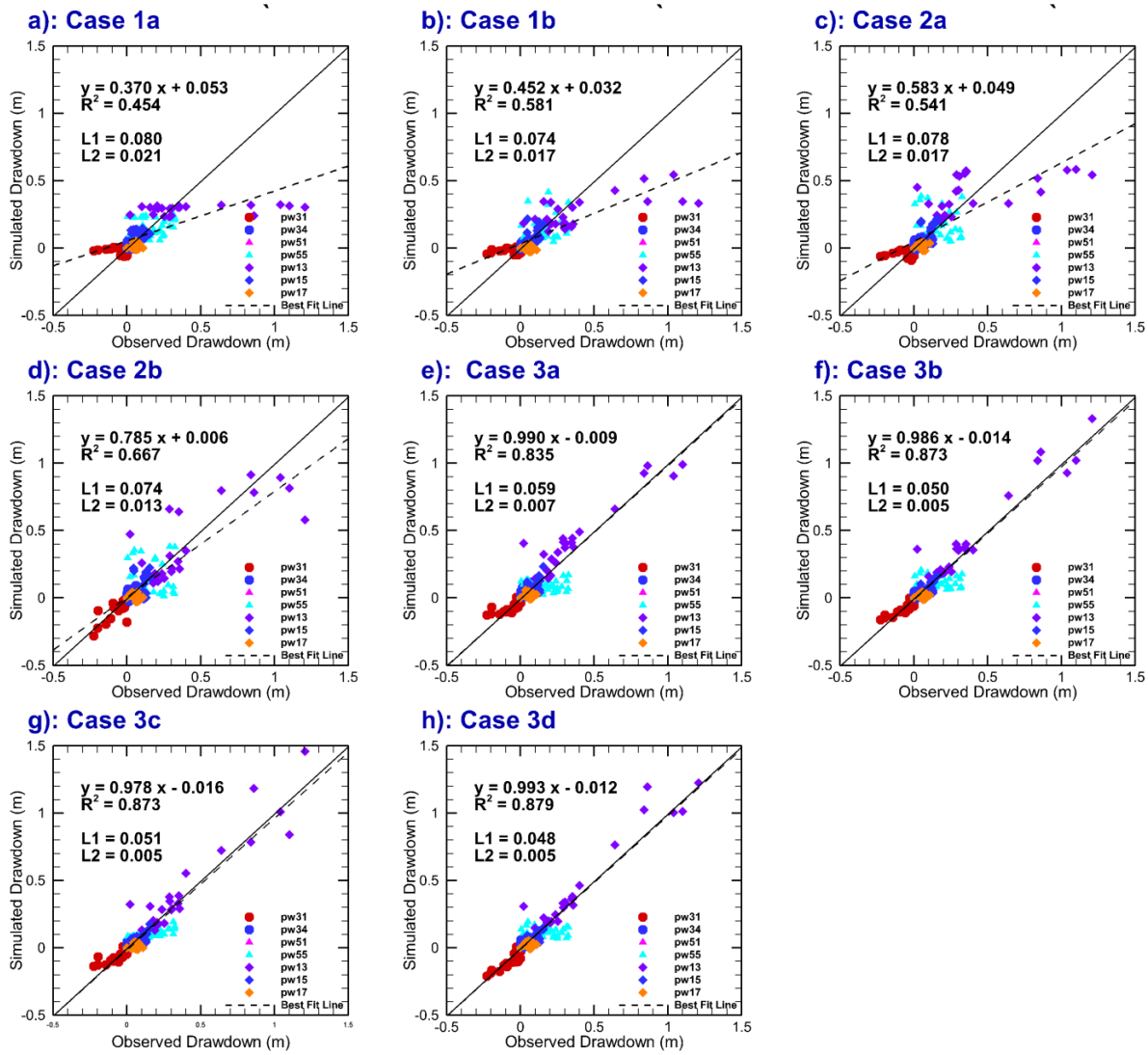


Figure 5.7: Scatterplots of observed versus simulated drawdowns for model validations using seven pumping tests for the: (a) isotropic effective parameter model; (b) anisotropic effective parameter model; (c) geological model with five layers; (d) geological model with 19 layers; (e) SimSLE starting with $K = 8.0 \times 10^{-6}$ m/s as prior mean; (f) SimSLE using the calibrated five-layer geological model as prior distribution; (g) SimSLE using the calibrated 19-layer geological model as prior distribution; and (h) SimSLE using the uncalibrated 19-layer geological model assigned with permeameter K values as prior distribution. The solid line is a 1:1 line indicating a perfect match. The dash line is the best fit line. The linear fit results are also included on each plot.

geological information.

On the other hand, the calibration of the 19-layer model (Case 2b) took more than 800 model calls, due to the increased number of unknowns and for the reasons discussed earlier. The linear model fit of the drawdown scatterplot showed an improvement over the 5-layer model, with an increased slope of 0.74, an increased R^2 to 0.76, and lowered L_1 and L_2 norms (Figure C2). Examination of Figure C2 in the Supplementary Information Section reveals that the calibration of the 5-layer geological model ranked 6^{th} for both L_1 and L_2, while the calibration of the 19-layer geological model ranked 5^{th} for both L_1 and L_2. The estimated K values and corresponding 95% confidence intervals are listed in Table 5.3 and the corresponding K distribution is presented in Figure 5.4b.

Examination of Table 5.3 reveals that estimated K values for layers 3 and 5 have significantly large 95% confidence intervals compared to those of the other layers. One reason for this is that layers 3 and 5 only exist in narrow portions of the geological model and also far from the pumping and observation wells, as shown in Figure 5.2, thus very few or no observation data are available in these layers to provide the pressure head information for model calibration. Similar results are found in *Zhao et al.* (2016) through their laboratory sandbox study where the geological zonation information is perfectly known.

Comparing the results in Table 5.3 to Figure 5.4a, K estimates for the main sand layers of the 19-layer model show some similarities to the 5-layer geological model, by estimating a relatively high K value for Layer 15, while estimating a low K value for Layer 11. Differences between the two geological models are obvious from Figure 5.4a and 5.4b. Firstly, more details about the interlayering of high and low permeability zones are revealed in Case 2b for the upper part of domain than in Case 2a. Secondly, variations of K estimates are introduced for low permeable layers (layer 16 to layer 19) at the bottom

Table 5.3: Soil type, permeameter test K, estimated K values and corresponding posterior 95% confidence interval limits for the 19-layer geological model.

Layer	Soil Type	Permeameter Test K (m/s)	Estimated K (m/s)	95% Confidence Intervals	
				Lower limit	Upper limit
1	Clay	2.65×10^{-7}	3.15×10^{-6}	2.88×10^{-7}	3.44×10^{-5}
2	Silt	7.76×10^{-7}	4.26×10^{-7}	5.64×10^{-15}	$3.21\times10^{+1}$
3	Sand	1.45×10^{-7}	5.83×10^{-7}	1.11×10^{-19}	$3.07\times10^{+6}$
4	Clay	6.09×10^{-8}	1.27×10^{-5}	1.32×10^{-7}	1.23×10^{-3}
5	Sand and Silt	2.38×10^{-6a}	8.51×10^{-8}	8.41×10^{-31}	$8.60\times10^{+15}$
6	sandy Silt	2.38×10^{-6}	3.53×10^{-8}	2.50×10^{-9}	4.97×10^{-7}
7	Silt	3.13×10^{-7}	4.01×10^{-8}	1.34×10^{-8}	1.20×10^{-7}
8	Clay	1.82×10^{-6}	1.35×10^{-5}	2.86×10^{-6}	6.36×10^{-5}
9	sandy Silt	5.04×10^{-6}	5.40×10^{-5}	2.14×10^{-5}	1.36×10^{-4}
10	Silt	7.47×10^{-6}	2.34×10^{-5}	4.50×10^{-6}	1.22×10^{-4}
11	Sand	1.32×10^{-6}	1.05×10^{-7}	4.90×10^{-8}	2.23×10^{-7}
12	Clay	3.74×10^{-7}	3.66×10^{-8}	5.54×10^{-9}	2.42×10^{-7}
13	sandy Silt	1.17×10^{-6}	6.29×10^{-5}	3.15×10^{-5}	1.25×10^{-4}
14	Silt	1.13×10^{-7}	6.27×10^{-7}	2.82×10^{-7}	1.39×10^{-6}
15	Sand and Gravel	1.22×10^{-5}	6.66×10^{-5}	5.31×10^{-5}	8.35×10^{-5}
16	Clay	2.01×10^{-8}	5.84×10^{-8}	2.32×10^{-9}	1.47×10^{-6}
17	Silt and Clay	2.44×10^{-8}	4.18×10^{-7}	2.95×10^{-11}	5.94×10^{-3}
18	Clay	4.72×10^{-9}	2.37×10^{-7}	6.16×10^{-13}	9.15×10^{-2}
19	Silt and Clay	2.44×10^{-8b}	1.70×10^{-5}	8.49×10^{-8}	3.41×10^{-3}

[a] K value for layer 5 is estimated as the value of layer 6;
[b] K value for layer 19 is estimated as the value of layer 17.

of the study area. More quantitative comparisons between the calibrated geological models with the permeameter test results will be provided later.

5.3.3 Case 3 Results

Four scenarios are considered for the geostatistical inversion approach using different prior distributions. The L_2 norm changes during the calibration process for all four scenarios and are plotted in the Supplementary Section as Figure C3. We selected inversion results from the iteration step at which the L_2 norm has stabilized indicating the convergence of the inversion process as suggested by *Xiang et al.* (2009). The result from the 82^{th} iteration is selected for Case 3a, while results from the 62^{th} iteration are selected for Cases 3b, 3c and 3d (Figure C3). It is interesting to note that Case 3a with a uniform K value takes more iterations to converge, when compared to Cases 3b – 3d in which various geological

models are used as prior distributions.

For Case 3a, a uniform mean K field is used as the prior distribution for the geostatistical inversion before SimSLE starts to iteratively condition the parameter field with pressure head measurements. The drawdown scatterplot of Case 3a (Figure 5.3e) shows significant improvement over the effective parameter (Case 1a and 1b) and the geological zonation (Cases 2a and 2b) approaches, in terms of the R^2 as well as L_1 and L_2 norms. However, we see that there is an obvious drift for the data from PW1-6 from the 45 degree line compared to those from the other pumping tests.

We observe from Figure C2 that all four geostatistical inversion scenarios yielded quite similar L_1 and L_2 norms at the selected iteration number and ranked in the top four consistently with the specified rankings given between the parentheses: Case 3a (4), Case 3b (3), Case 3c (1), and Case 3d (2).

Figure 5.4c provides the estimated K tomogram for Case 3a, while its corresponding residual variance of ln K in Figure 5.5a. Examination of Figure 5.4c reveals that, in general, the interlayering patterns of the high and low permeable zones are captured in the central part of modeling domain, where ln K residual variances (Figure 5.5a) are lower. In addition, a higher K zone is visible on the bottom left portion of the domain.

In the previous work of *Berg and Illman* (2013) who utilized four pumping tests (PW4-3, PW1-3, PW5-3 and PW3-3) to conduct SSHT analysis of the same area, the entire bottom area of the central model domain was estimated to have high K values, despite the fact that core samples indicated the presence of low K silt and clay layers. In addition, the lowermost ports situated in the low K zones did not yield measurable drawdown responses during those tests.

In contrast, for this study, we obtained measurable drawdown responses from the bot-

tom observation ports by pumping from port PW1-6 (Figure 5.1b) located in the lower part of the domain. By including these additional drawdown data from the low permeable zone, the inversion of all tests yielded slightly improved K estimates without showing the entire bottom area as a high K zone (*Berg and Illman*, 2011b, 2013). However, our results (Figure 5.4a) are still inconsistent with the known geology consisting of silt and clay layers (Figure 5.2). For example, in the bottom left portion of the domain beyond the central 15 m by 15 m well cluster area, the K estimates (Figure 5.4c) and the residual variances of ln K (Figure 5.5a) are generally high, due to the fact that no wells and pressure head data are available in that region for model calibration.

In Case 3b, we extended Case 3a by using the K tomogram obtained from Case 2a, which is the calibrated 5-layer geological model as the prior distribution for the geostatistical inverse model. The drawdown scatterplot of Case 3b (Figure 5.3f) reveals an obvious improvement compared to Case 3a (Figure 5.3e), in which a uniform mean K field is used as the prior distribution. On the other hand, obvious differences can be seen in the estimated K tomogram from Case 3b (Figure 5.4d) when compared to Case 3a (Figure 5.4c). Specifically, K estimates from Case 3b (Figure 5.4d) reveal a pattern that preserves the geological features of the K distribution of the calibrated 5-layer model, as well as the heterogeneity features in the upper part of the K tomogram for Case 3a (Figure 5.4c).

For the bottom part of the simulation domain, the estimated K values in Case 3b (Figure 5.4d) are significantly lower than Case 3a (Figure 5.4c). In addition, the low K zone at the bottom of the simulation domain extends across the site. Both of these features in Case 3d are more consistent with our knowledge of site geology (Figure 5.2).

The residual variance of ln K for Case 3b (Figure 5.5b) reveals that the variances are relatively low within and in the vicinity of the well field. However, as in Case 3a, the

variances are higher away from the well field.

The geostatistical inverse modeling of Case 3c was performed by using the K estimates of the calibrated 19-layer geological model of Case 2b as the prior distribution. From Figure 5.3, we observe that the calibration scatterplot between simulated and observed drawdowns for Case 3c (Figure 5.3g) shows slight improvements compared to Cases 3a (Figure 5.3e) and 3b (Figure 5.3f). The estimated K tomogram and the corresponding ln K residual variance are provided in Figures 5.4e and 5.5c, respectively. Generally, the main layering pattern shown in Figure 5.4e follows the pattern of the calibrated 19-layer geological model (Figure 5.4b). In addition, we can clearly see more details to the geological features throughout the site, because a 19-layer model is used as the prior distribution.

Case 3d uses the 19-layer geological model populated with permeameter test K values as a prior distribution for geostatistical inverse modeling. This case could be viewed as the scenario with most data included into the inverse model among all four geostatistical inversion cases including pressure heads, geological data, and local K data from permeameter tests.

Figure 5.3h provides the drawdown scatterplot for Case 3d which shows that the R^2 value as well as the L_1 and L_2 norms have improved over Case 3a. However, the results are comparable to Cases 3b (Figure 5.3f) and 3c (Figure 5.3g) suggesting that including permeameter K data as prior information has not significantly improved the calibration results.

Figure 5.4f provides the estimated K tomogram from Case 3d and the corresponding ln K variance in Figure 5.5d. Compared to the K tomogram for Case 3a (Figure 5.4c), the structural features shown in the geological model (Figure 5.2) are better preserved in the recovered K tomogram (Figure 5.4f). Similar to Cases 3b (Figure 5.4d) and 3c (Figure

5.4e), the heterogeneous K distributions and in particular, the layering is similar to prior values. However, due to the inclusion of permeameter K data into the Case 3d model, the K values for the lowermost layer consisting of silt and clay are more representative of site geology in comparison to Cases 3a (Figure 5.4c), 3b (Figure 5.4d), and 3c (Figure 5.4e). The residual variance map of ln K (Figure 5d) is similar to the other cases (Figures 5a – 5c).

Results obtained from calibrating Cases 3b, 3c and 3d suggest that when geologically distributed K fields are used as prior distributions, HT analysis using the geostatistical inversion approach yields K tomograms with geological features. This would be helpful for HT to correctly capture the stratigraphic features for areas where only limited pressure head data are available.

5.4 Comparison of estimated K with Permeameter Test K

We then compared the estimated K values of all scenarios from Cases 2 and 3 to permeameter K values obtained along the CMT and PW wells, as shown in Figure 5.6. This comparison enabled us to examine the performance of different subsurface modeling approaches in terms of both intra- and interlayer K variations. Specifically, the calibration of the simple 5-layer geological model in Case 2a captured the variation trend of vertical K within a range of 0 m to 10 m above the bottom of modeling domain, while the K variation in the 10 m to 15 m zone were generally missed due to the merging of layers in the simplified geological model. It is important to keep in mind that once the geological model is constructed, we did not make adjustments to the geological structure during the

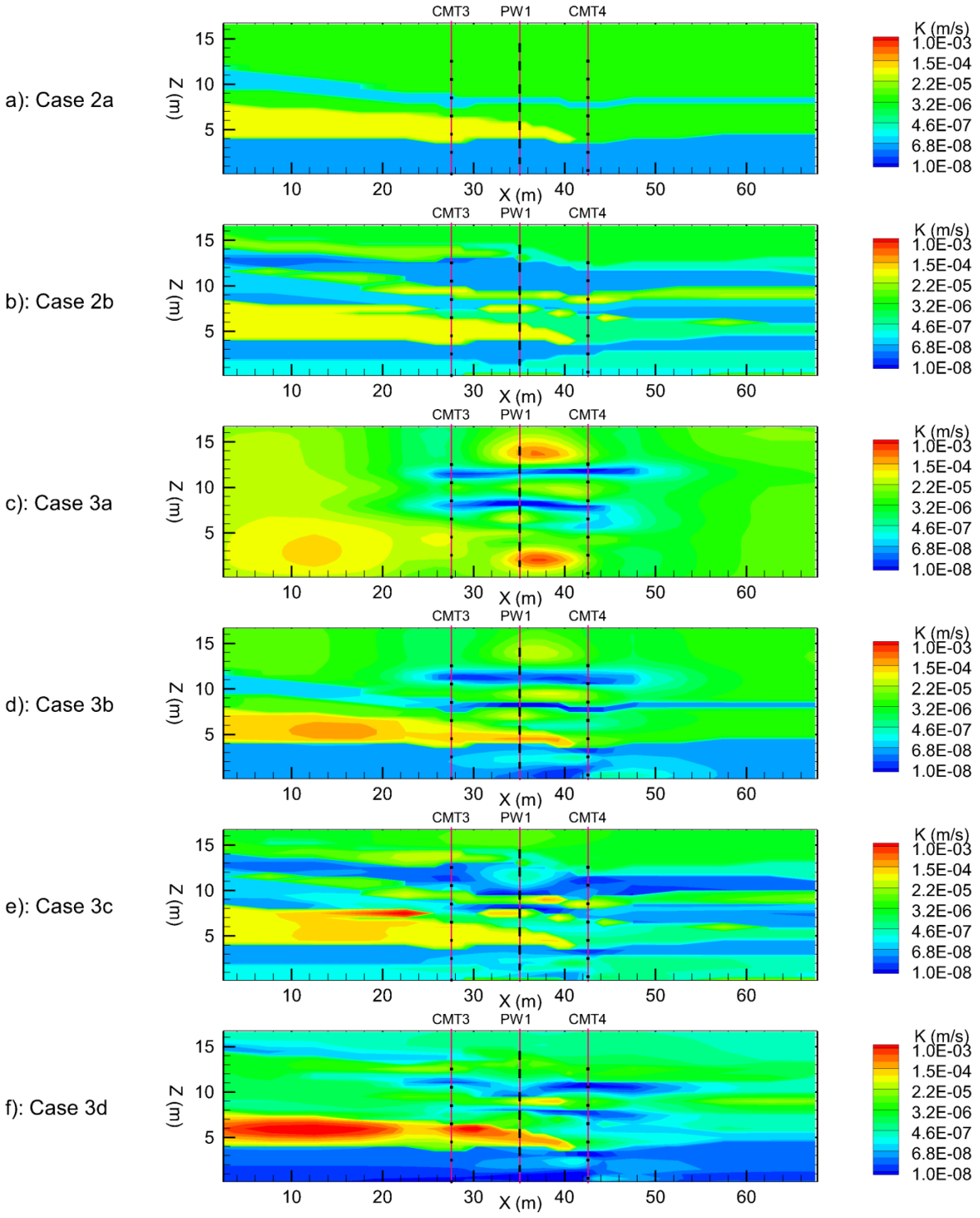


Figure 5.8: K estimates along *D-D'* cross section for the: (a) 5-layer geological model; (b) 19-layer geological model; (c) SimSLE starting with $K = 8.0 \times 10^{-6}$ m/s as prior mean; (d) SimSLE using the calibrated five-layer geological model as prior distribution; (e) SimSLE using the calibrated 19-layer geological model as prior distribution; (f) SimSLE using the uncalibrated 19-layer geological model assigned with permeameter K values as prior distribution.

calibration phase of the study, although it is conceivable that such adjustments could be made. We have not made such adjustments for this study because such modifications are typically not attempted. Instead, the parameters of each geologic unit are adjusted during the calibration phase.

In contrast, by treating the K values of each layer separately as unknowns, the 19-layer geological model of Case 2b yields refined fitting of K profiles for the top 5 m to 15 m zone of the simulation domain, which can be clearly seen from Figure 5.6a and 5.6b, due to the larger number of adjustable parameters (19 for the model Geo-19) for calibration. Meanwhile, K profiles obtained from both calibrated geological models showed some inconsistency to permeameter K data along nine wells. Such inconsistency could be attributed, on one hand, to using geological zonations with each layer to be homogeneous, and on the other hand, to the compensation effect of parameter values to structural errors (*Refsgaard et al.*, 2012). These results collectively suggest that calibration of geological models interpolated from borehole logs to multiple pumping tests is useful in terms of providing general K estimates of the field. However, because the stratigraphy of geological models is fixed in this study, fine scale variability in K within each layer cannot be captured. Next, we compare the estimated K values from the geostatistical inversion approaches (Cases 3a – 3d) to the permeameter K data.

A comparison of the results from Case 3a to permeameter test K values reveals that, when a homogeneous K field is used as the prior mean, the geostatistical inversion approach only captures the general features of high and low permeable layers within the range of 5 m to 12 m, and K estimates for the area away from the well field is relatively smooth. The main reason for this is that no observation data are available to update the K estimates during SimSLE inversion (*Xiang et al.*, 2009). However, as shown in Figure 5.6, when

geologically distributed K fields are used as prior distributions (Case 3b, Case 3c and Case 3d), the fits between the estimated and permeameter tested K values of all CMT wells and PW wells are consistently improved. The improvements are most obvious for the high K zone located between 4 m to 7 m above the bottom of the modeling domain, as well as the low K zone near the bottom. This result supports the finding by **?** through a synthetic case study, that the prior knowledge of the site-specific geological structures can be important for resolving the correct aquifer/aquitard features.

Additionally, we find from Figure 6 that the fit of K profiles in Case 3b with a 5-layer geological model used as a prior distribution in the geostatistical inversion approach is comparable to those from Cases 3c and 3d, in which a 19-layer geological model is used. This finding indicates that a simple geological model reflecting the general geological structure may be sufficient for being used as prior distribution in geostatistical inversion approaches to characterize heterogeneity within the area of interest.

Another important feature of the estimated K tomogram of Case 3a (Figure 5.4c) is the incorrect mapping of the clay zone at the bottom of the simulation domain, which is the same finding by the previous studies of *Berg and Illman* (2011b, 2013). This is so despite that additional steady state drawdown data from the pumping tests at PW1-6, where the pumping took place in the bottommost low K zone was included for inverse modeling.

We see from the calibration scatterplot Figure 5.3e that fitting of pumping test PW1-6 is poorer than the other tests. A potential cause for this poorer fit is that insufficient amount of steady state drawdown information was provided to constrain the inversion process and to guide the inverse model to correctly estimate more representative values. However, the simple, five-layer geological model correctly identifies the lowermost layer as a low K zone. One possible explanation is that the lower complexity 5-layer geological

model is more likely to be justified than the highly parameterized complex geostatistical model (*Schöniger et al.*, 2015), with the given amount of observed pressure head data from the lower portion of the simulation domain.

One potential solution to resolve the issue of inconsistency with geological knowledge of the site may be to properly regularize the inverse problem as suggested by *Doherty* (2003) and *Carrera et al.* (2005). Additionally, a THT analysis of the same pumping tests may improve these results, since we are able to include more pressure head data to further constrain the inverse problem. Such a study is currently under progress and will be reported in the future.

Overall, the above comparisons suggest that the use of geological data is helpful for the geostatistical inversion approach for HT investigations, in preserving structural features of the hydraulic property field.

5.5 Prediction of Steady State Drawdowns

Next, we use the estimated K fields from all cases to predict the drawdowns of seven additional pumping tests (PW3-1, PW 3-4, PW5-5, PW1-3, PW1-5 and PW1-7) not used in the inverse modeling effort. These tests are selected from pumping ports located at different areas of the modeling domain. For each pumping test, we selected only the drawdown curves of the observation ports that reached steady state. The scatterplots of observed and simulated drawdowns are shown in Figure 7. Linear fit results and L_1 and L_2 norms are also included to evaluate the overall prediction performance for the seven pumping tests. Meanwhile, Figure C4 in the Supplementary Information section summarizes the L_1 and L_2 norms from each pumping test and their performance ranking of drawdown predictions.

Examinations of Figures 5.7a-5.7h and C4 reveal that, the results of drawdown pre-

dictions improve gradually from the effective parameter approach (Case 1) to the highly parameterized approach based on geostatistical inverse modeling (Case 3). The isotropic effective parameter model Case 1a yielded results that ranked at the bottom in terms of both L_1 and L_2 norms for drawdown prediction, while considering anisotropy in Case 1b (6^{th}: L_1 norm and 7^{th}: L_2 norm) or using a 5-layer geological model (7^{th}: L_1 norm and 6^{th}: L_2 norm) improved the results. Additionally, prediction results of the complex 19-layer model geological model ranked in the middle (5^{th} in terms of L_1 and L_2). The geostatistical model ranked at the top four in terms of L_1 and L_2 norms with very close prediction performances. Specifically, geostatistical inversion Case 3a using a uniform prior mean ranked 4^{th}, geostatistical inversion Case 3b using the 5-layer calibrated geological model as prior distribution ranked 2^{nd} for both L_1 and L_2, Case 3c using the calibrated 19-layer geological model as prior distribution ranked 3^{rd}, and Case 3d using the uncalibrated geological model populated with permeameter K data as prior distribution ranked the 1^{st}. These results are consistent with findings from the laboratory sandbox study of *Illman et al.* (2015), that the geostatistical inversion approach performed the best in terms of drawdown predictions when compared with effective parameter and geological modelling approaches.

When geologically distributed K values were used as prior distributions, it is interesting to note that the geostatistical inversion Cases 3b, 3c and 3d performed quite closely in terms of model calibration and validation and only slightly better in terms of R^2, L_1 and L_2 norms than Case 3a, in which an uniform K prior mean value was used, given the differences in the estimated K tomograms (Figures 5.4c - 5.4f). To examine this issue, we plotted K estimates for geological and geostatistical models along cross section *D-D'* for a detailed comparison (shown in Figure 5.8). Examination of the cross-sections reveals that there are large differences in the morphology of zones.

These results on one hand, suggest that prior K distributions could affect the inversion results differently when a highly parameterized approach is utilized in HT investigations. Thus, one must use caution in selecting reasonable values for prior means. On the other hand, we want to restate that the above comparisons of drawdown predictions were based only on data from observation ports that reached steady state during each pumping test. Most of the selected observation ports during each pumping test were located in the upper and central part of the simulation domain. The number of observation ports located in the lower part of the modeling domain selected for validation purposes was much lower, due to the fact that no obvious responses were observed during pumping from ports in the upper and central part of the domain (PW3-1, PW5-1, PW1-3, and PW 3-4) and that the steady state had not been reached for model validation pumping tests from the lower part of the domain (PW1-5 and 1-7). However, available drawdown curves indicate responses in the observation ports located in the lowermost portion of the simulation domain. Thus, THT studies will be needed at this site to provide more complete evaluations of drawdown prediction performances between geological and geostatistical inversion approaches.

Conclusions and Recommendations

6.1 Conclusions

Conclusions of this thesis are grouped in three parts.

Part I: Hydraulic Tomography in Unconfined Aquifer: Sandbox Study

- Based on the results of laboratory sandbox experiments, HT analysis that considers variably saturated flow processes is able to vividly characterize the heterogeneous distribution of K in both saturated and unsaturated zones. When unsaturated flow is also considered, HT could capture the main heterogeneity pattern of the soil parameters for the unsaturated zone, such as the α and θ_s of the Gardner-Russo's model.

- HT inverted K tomogram yields significantly improved drawdown predictions compared to models based on the homogeneous assumption, especially at early and intermediate time.

- When heterogeneity of unsaturated zone parameters is ignored, HT inverted K tomogram loses some accuracy compared to the case that considers both heterogeneous saturated and unsaturated zone parameters. Thus, the HT algorithm that considers variably saturated flow should be favored for accurate and higher-resolution characterization of site heterogeneity for unconfined aquifers.

Part II: On the Importance of Geological Data in Hydraulic tomography: Sandbox Study

The results of steady state HT study in a laboratory sandbox lead to the following findings and conclusions:

- When HT test data are jointly calibrated for layer-based geological models, geological models that more closely represented the aquifer stratigraphy performs better than geological models with errors in stratigraphy in terms of both calibration and prediction of drawdowns.

- This sandbox study also shows that even poor geological models could be calibrated quite well, while K estimates for each unit can be unreasonable compared to the permeameter K measurements due to the compensational effect of estimated parameters for model structure error (Refsgaard et al., 2012 and Illman et al., 2015).

- When abundant HT data are used, the geostatistical inverse modeling approach performs better than both good and poor geological models constructed from borehole logs in terms of model calibration and validation. When the amount of data used for calibration are reduced, performance gaps are reduced between the geostatistical inversion model and the geological model that closely represents the aquifer.

- The use of geological models as prior K values for the geostatistical inversion approach leads to only slight improvements in model calibration and validation results, however, leads to the preservation of geological structures.

Part III: On the Importance of Geological Data on Hydraulic tomography: NCRS Field Study

- Simultaneous calibration of geological models to multiple pumping test data yields K estimates that generally reflect the K variation trends of permeameter test values along nine boreholes for the highly heterogeneous, glaciofluvial aquifer-aquitard system at NCRS.

- Geological data is helpful in improving the results of HT when hydraulic test data are lacking through its use as a prior distribution in the geostatistical inversion approach. Specifically, using reliable geological models instead of an effective K value as prior mean fields is helpful for the geostatistical inverse approach in improving K correspondence with permeameter test results and in preserving of geological features for K tomograms.

6.2 Practical Recommendations and Future Studies

6.2.1 Practical Recommendations

The past HT studies have shown the importance of accurate delineating the hydraulic heterogeneity for flow predictions and solute transport problems *Illman et al.* (2010b, 2012), mostly at laboratory and small field-site scales. Until now, only *Illman et al.* (2009) and *Zha et al.* (2015, 2016) have only present studies of hydraulic tomography conducted at the kilometer scale at fractured rock site. Thus, (*Yeh et al.*, 2009) has proposed a river stage tomography approach which takes advantage of natural stimuli such as flow changes in river stages to estimate the aquifer properties over the basin scale.

Practically speaking, field implications of HT approach relying solely on pumping test data could be limited due to the several factors, such as the domain size of the detectable signals, duration of the the pumping test, and the quality of drawdown signals (*Illman*,

2014). For future practical applications, HT could be helpful at field sites where high resolution characterizations are needed and pumping data exists, such as at remediation sites with pump and treat operations or where chemicals are injected into the subsurface to destroy contaminants in the source zones.

Overall, HT should be useful where more accurate maps of subsurface hydraulic parameter heterogeneity are needed and its uncertainty quantified. The main strength of HT is the accurate mapping of the hydraulic parameter distributions. As shown in this thesis and previous studies, model calibration and validation performances reveal the robustness of HT results. Another important strength of HT surveys over traditional subsurface characterization methods is the accurate mapping of continuity/connectivity of high or low K features. Such features often dictate groundwater flow and its storage. In terms of contaminant transport, connected high K features can allow rapid migration of contaminants. Alternatively, low K features can act as long term storage reservoirs of contaminants. Therefore, the accurate delineation of these features are essential in managing groundwater resources and to clean them up efficiently when they become contaminated.

Based on the findings and conclusions of this thesis, some practical guidance can be provided as following:

- In Part I of this thesis, while the setup of the experiments may not faithfully represent the setup of pumping tests in unconfined aquifers under field conditions, the study is an intermediate step for its future applications to field aquifers. The results emphasize the necessities of considering the variably saturated flow and unsaturated zone heterogeneity during site characterization of the unconfined aquifers.

- Conclusions of studies in Part II and Part III show the importance of using geological models for steady state HT. That is, reliable geological information and pressure head

data should be jointly used for HT in order to produce refinements of the recovered parameter distributions, which could in turn lead to better flow predictions.

6.2.2 Future Studies

Although HT has been developed and extensively tested from synthetic to field studies for almost two decades, there are still many questions that remain unanswered and need to be investigated for future research, such as:

- Evaluating the performances of geological models for HT analysis under transient flow conditions for both laboratory and field conditions;
- Developing new geostatistical inversion methods for HT to directly take advantage of different types of information (e.g., flux measurement, borehole logs, ground penetrating radar data) other than pressure heads;
- Considering K anisotropy in geostatistical inversion methods for HT analysis for highly heterogeneous sites.

Bibliography

M. Alexander, S. J. Berg, and W. A. Illman. (2011). Field Study of Hydrogeologic Characterization Methods in a Heterogeneous Aquifer. *Ground Water*, 49(3):365–382. ISSN 0017467X. doi: 10.1111/j.1745-6584.2010.00729.x.

S. J. Berg and W. A. Illman. (2011a). Capturing aquifer heterogeneity: Comparison of approaches through controlled sandbox experiments. *Water Resources Research*, 47(9): n/a–n/a. ISSN 1944-7973. doi: 10.1029/2011WR010429. W09514.

S. J. Berg and W. A. Illman. (2011b). Three-dimensional transient hydraulic tomography in a highly heterogeneous glaciofluvial aquifer-aquitard system. *Water Resources Research*, 47(10). ISSN 00431397. doi: 10.1029/2011WR010616.

S. J. Berg and W. A. Illman. (2012). Improved predictions of saturated and unsaturated zone drawdowns in a heterogeneous unconfined aquifer via transient hydraulic tomography: Laboratory sandbox experiments. *Journal of Hydrology*, 470-471:172–183. ISSN 00221694. doi: 10.1016/j.jhydrol.2012.08.044. URL `http://dx.doi.org/10.1016/j.jhydrol.2012.08.044`.

S. J. Berg and W. A. Illman. (2013). Field study of subsurface heterogeneity with steady-state hydraulic tomography. *Ground Water*, 51(1):29–40. ISSN 0017467X. doi: 10.1111/j.1745-6584.2012.00914.x.

S. J. Berg and W. A. Illman. (2015). Comparison of Hydraulic Tomography with Traditional Methods at a Highly Heterogeneous Site. *Ground Water*, 53(1):71–89. ISSN 17456584. doi: 10.1111/gwat.12159.

S. J. Berg, P. A. Hsieh, and W. A. Illman. (2011). Estimating Hydraulic Parameters When Poroelastic Effects Are Significant. *Ground Water*, 49(6):815–829. ISSN 0017467X. doi: 10.1111/j.1745-6584.2010.00781.x.

M. Blouin, R. Martel, and E. Gloaguen. (2013). Accounting for Aquifer Heterogeneity from Geological Data to Management Tools. *Ground Water*, 51(3):421–431. ISSN 0017467X. doi: 10.1111/j.1745-6584.2012.00982.x.

J. M. Boggs, S. C. Young, L. M. Beard, L. W. Gelhar, K. R. Rehfeldt, and E. E. Adams. (dec 1992). Field study of dispersion in a heterogeneous aquifer: 1. Overview and site description. *Water Resources Research*, 28(12):3281–3291. ISSN 1944-7973. doi: 10.1029/92WR01756. URL `http://dx.doi.org/10.1029/92WR01756`.

G. C. Bohling and J. J. Butler. (2010). Inherent Limitations of Hydraulic Tomography. *Ground Water*, 48(6):809–824. ISSN 0017467X. doi: 10.1111/j.1745-6584.2010.00757.x.

G. C. Bohling, X. Zhan, J. J. Butler, and L. Zheng. (2002). Steady shape analysis of tomographic pumping tests for characterization of aquifer heterogeneities. *Water Resources Research*, 38(12):60–1–60–15. ISSN 1944-7973. doi: 10.1029/2001WR001176. 1324.

G. C. Bohling, X. Zhan, M. D. Knoll, J. J. Butler, and B. G. C. Bohling. (2003). Hydraulic Tomography and the Impact of A Priori Information: An Alluvial Aquifer Example. pages 1–16. URL `http://www.kgs.ku.edu/Hydro/WellTests/OFR03{_}71/OFR03-71.pdf`.

G. C. Bohling, J. J. Butler, X. Zhan, and M. D. Knoll. (2007). A field assessment of the value of steady shape hydraulic tomography for characterization of aquifer heterogeneities. *Water Resources Research*, 43(5):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2006WR004932. W05430.

N. S. Boulton. (1954). Unsteady radial flow to a pumped well allowing for delayed yield from storage. *International Association of Hydrological Sciences*, 2:472–477.

N. S. Boulton. (1963). Analysis of data from pumping tests allowing for delayed yield from storage. *Proceedings of the Institution of Civil Engineers*, 26(3):469–482. doi: 10.1680/iicep.1963.10409.

R. Brauchler, R. Liedl, and P. Dietrich. (2003). A travel time based hydraulic tomographic approach. *Water Resour. Res.*, 39(12):1370. ISSN 00431397. doi: 10.1029/2003WR002262. URL `http://doi.wiley.com/10.1029/2003WR002262`.

R. Brauchler, J.-T. Cheng, P. Dietrich, M. Everett, B. Johnson, R. Liedl, and M. Sauter. (2007). An inversion strategy for hydraulic tomography: Coupling travel time and amplitude inversion. *Journal of Hydrology*, 345(3–4):184–198. ISSN 0022-1694. doi: http://dx.doi.org/10.1016/j.jhydrol.2007.08.011. URL `http://www.sciencedirect.com/science/article/pii/S0022169407004428`.

R. Brauchler, R. Hu, T. Vogt, D. Al-Halbouni, T. Heinrichs, T. Ptak, and M. Sauter. (2010). Cross-well slug interference tests: An effective characterization method for resolving aquifer heterogeneity. *Journal of Hydrology*, 384(1–2):33–45. ISSN 0022-1694. doi: http://dx.doi.org/10.1016/j.jhydrol.2010.01.004. URL `http://www.sciencedirect.com/science/article/pii/S0022169410000156`.

R. Brauchler, R. Hu, P. Dietrich, and M. Sauter. (2011). A field assessment of high-resolution aquifer characterization based on hydraulic travel time and hydraulic attenuation tomography. *Water Resources Research*, 47(3):1–12. ISSN 00431397. doi: 10.1029/2010WR009635.

R. Brauchler, J. Doetsch, P. Dietrich, and M. Sauter. (2012). Derivation of site-specific relationships between hydraulic parameters and p-wave velocities based on hydraulic and seismic tomography. *Water Resources Research*, 48(3):1–14. ISSN 00431397. doi: 10.1029/2011WR010868.

M. Cardiff and W. Barrash. (2011). 3-d transient hydraulic tomography in unconfined aquifers with fast drainage response. *Water Resources Research*, 47(12):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2010WR010367. W12518.

M. Cardiff, W. Barrash, P. Kitanidis, B. Malama, A. Revil, S. Straface, and E. Rizzo. (2009). A potential-based inversion of unconfined steady-state hydraulic tomography. *Ground Water*, 47(2):259–270. ISSN 1745-6584. doi: 10.1111/j.1745-6584.2008.00541.x.

M. Cardiff, W. Barrash, and P. K. Kitanidis. (2012). A field proof-of-concept of aquifer imaging using 3-d transient hydraulic tomography with modular, temporarily-emplaced equipment. *Water Resources Research*, 48(5):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2011WR011704. W05531.

M. Cardiff, W. Barrash, and P. K. Kitanidis. (2013a). Hydraulic conductivity imaging from 3-d transient hydraulic tomography at several pumping/observation densities. *Water Resources Research*, 49(11):7311–7326. ISSN 1944-7973. doi: 10.1002/wrcr.20519.

M. Cardiff, T. Bakhos, P. K. Kitanidis, and W. Barrash. (2013b). Aquifer heterogeneity characterization with oscillatory pumping: Sensitivity analysis and imaging potential. *Water Resources Research*, 49(9):5395–5410. ISSN 1944-7973. doi: 10.1002/wrcr.20356.

S. Carle and G. Fogg. (1997). Modeling Spatial Variability with One and Multidimensional Continuous-Lag Markov Chains. *Mathematical Geology*, 29(7):891–918. ISSN 0882-8121. doi: 10.1023/a:1022303706942.

J. Carrera, A. Alcolea, A. Medina, J. Hidalgo, and L. J. Slooten. (2005). Inverse problem in hydrogeology. *Hydrogeology Journal*, 13(1):206–222. ISSN 1435-0157. doi: 10.1007/s10040-004-0404-7. URL `http://dx.doi.org/10.1007/s10040-004-0404-7`.

M. Castagna and A. Bellin. (2009). A Bayesian approach for inversion of hydraulic tomographic data. *Water Resources Research*, 45(4):1–16. ISSN 00431397. doi: 10.1029/2008WR007078.

M. Castagna, M. W. Becker, and A. Bellin. (2011). Joint estimation of transmissivity and storativity in a bedrock fracture. *Water Resources Research*, 47(9):1–19. ISSN 00431397. doi: 10.1029/2010WR009262.

W. P. Clement, W. Barrash, and M. D. Knoll. (may 2006). Reflectivity modeling of a ground-penetrating-radar profile of a saturated fluvial formation. *GEOPHYSICS*, 71(3): K59–K66. ISSN 0016-8033. doi: 10.1190/1.2194528. URL `http://library.seg.org/doi/10.1190/1.2194528`.

A. Comunian, P. Renard, J. Straubhaar, and P. Bayer. (2011). Three-dimensional high resolution fluvio-glacial aquifer analog - Part 2: Geostatistical modeling. *Journal of Hydrology*, 405(1-2):10–23. ISSN 00221694. doi: 10.1016/j.jhydrol.2011.03.037.

A. J. Craig. (2005). *Measurement of hydraulic parameters at multiple scales in two synthetic heterogeneous aquifers constructed in the laboratory.* M.s. thesis, Dept. of Civ. and Environ. Eng., Univ. of Iowa, Iowa City.

G. de Marsily, F. Delay, J. Gonçalvès, P. Renard, V. Teles, and S. Violette. (2005). Dealing with spatial heterogeneity. *Hydrogeology Journal*, 13(1):161–183. ISSN 1435-0157. doi: 10.1007/s10040-004-0432-3.

J. Doherty. (1994). PEST: Model-Independent Parameter Estimation.

J. Doherty. (mar 2003). Ground Water Model Calibration Using Pilot Points and Regularization. *Ground Water*, 41(2):170–177. ISSN 1745-6584. doi: 10.1111/j.1745-6584.2003.tb02580.x. URL `http://dx.doi.org/10.1111/j.1745-6584.2003.tb02580.x`.

J. Doherty. (2005). PEST: Model-Independent Parameter Estimation User Manual, 5th ed., Watermark Numer. Comput., Brisbane, Australia.

A. L. Endres, J. P. Jones, and E. A. Bertrand. (2007). Pumping-induced vadose zone drainage and storage in an unconfined aquifer: A comparison of analytical model predictions and field measurements. *Journal of Hydrology*, 335(1–2):207 – 218. ISSN 0022-1694. doi: http://dx.doi.org/10.1016/j.jhydrol.2006.07.018.

R. N. Farvolden, P. Greenhouse, P. Karrow, P. Pehme, and L. Ross. (1987). Subsurface quaternary stratigraphy of the Kitchener-Waterloo area using borehole geophysics. Technical report, Toronto, Ontario:Ontario Geological Survey.

W. R. Gardner. (1958). Some steady state solutions of unsaturated moisture flow equations with applications to evaporation from a water table. *Soil Science*, 85(4). ISSN 0038-075X.

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

W. G. Harrar, T. O. Sonnenborg, and H. J. Henriksen. (2003). Capture zone, travel time, and solute-transport predictions using inverse modeling and different geological models. *Hydrogeology Journal*, 11(5):536–548. ISSN 14312174. doi: 10.1007/s10040-003-0276-2.

S. Y. Huang, J. C. Wen, T. C. J. Yeh, W. Lu, H. L. Juan, C. M. Tseng, J. H. Lee, and K. C. Chang. (2011). Robustness of joint interpretation of sequential pumping tests: Numerical and field experiments. *Water Resources Research*, 47(10):1–18. ISSN 00431397. doi: 10.1029/2011WR010698.

W. A. Illman. (2014). Hydraulic tomography offers improved imaging of heterogeneity in fractured rocks. *Groundwater*, 52(5):659–684. ISSN 17456584. doi: 10.1111/gwat.12119.

W. A. Illman, X. Liu, and A. Craig. (2007). Steady-state hydraulic tomography in a laboratory aquifer with deterministic heterogeneity: Multi-method and multiscale validation of hydraulic conductivity tomograms. *Journal of Hydrology*, 341(3–4):222 – 234. ISSN 0022-1694. doi: http://dx.doi.org/10.1016/j.jhydrol.2007.05.011.

W. a. Illman, X. Liu, and A. Craig. (2008). Evaluation of transient hydraulic tomography and common hydraulic characterization approaches through laboratory sandbox experiments. *Journal of Environmental Engineering and Management*, 18(4):249–256.

W. A. Illman, X. Liu, S. Takeuchi, T.-C. J. Yeh, K. Ando, and H. Saegusa. (2009). Hydraulic tomography in fractured granite: Mizunami underground research site, japan. *Water Resources Research*, 45(1):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2007WR006715. W01406.

W. A. Illman, J. Zhu, A. J. Craig, and D. Yin. (2010a). Comparison of aquifer characterization approaches through steady state groundwater model validation: A controlled

laboratory sandbox study. *Water Resources Research*, 46(4):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2009WR007745. W04502.

W. A. Illman, S. J. Berg, X. Liu, and A. Massi. (2010b). Hydraulic/partitioning tracer tomography for dnapl source zone characterization: Small-scale sandbox experiments. *Environmental Science and Technology*, 44(22):8609–8614. ISSN 0013936X. doi: 10.1021/es101654j.

W. A. Illman, S. J. Berg, and T.-C. J. Yeh. (2012). Comparison of approaches for predicting solute transport: Sandbox experiments. *Ground Water*, 50(3):421–431. ISSN 1745-6584. doi: 10.1111/j.1745-6584.2011.00859.x.

W. A. Illman, S. J. Berg, and Z. Zhao. (2015). Should hydraulic tomography data be interpreted using geostatistical inverse modeling? A laboratory sandbox investigation. *Water Resources Research*, 51(5):3219–3237. ISSN 00431397. doi: 10.1002/2014WR016552. URL `http://doi.wiley.com/10.1002/2014WR016552`.

A. G. Journel and J. J. Gomez-Hernandez. (1993). Stochastic Imaging of the Wilmington Clastic Sequence. *Spe*, (March):33–40. ISSN 0885923X.

A. G. Journel and E. H. Isaaks. (1984). Conditional indicator simulation: Application to a saskatchewan uranium deposit. *Journal of the International Association for Mathematical Geology*, 16(7):685–718. ISSN 1573-8868. doi: 10.1007/BF01033030.

P. Karrow. (1979). Geology of the University of Waterloo Campus. Technical report, Department of Earth Sciences, University of Waterloo, Waterloo, Ontario, Canada.

P. F. Karrow. (1993). Quaternary Geology, Stratford-Conestogo Area. Note, Ontario, Canada.

C. E. Koltermann and S. M. Gorelick. (1996). Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches. *Water Resources Research*, 32(9):2617–2658. ISSN 1944-7973. doi: 10.1029/96WR00025.

S. Liu, T.-C. J. Yeh, and R. Gardiner. (2002). Effectiveness of hydraulic tomography: Sandbox experiments. *Water Resources Research*, 38(4):5–1–5–9. ISSN 1944-7973. doi: 10.1029/2001WR000338.

X. Liu and P. K. Kitanidis. (2011). Large-scale inverse modeling with an application in hydraulic tomography. *Water Resources Research*, 47(2):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2010WR009144. W02501.

X. Liu, W. A. Illman, A. J. Craig, J. Zhu, and T.-C. J. Yeh. (2007). Laboratory sandbox validation of transient hydraulic tomography. *Water Resources Research*, 43(5):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2006WR005144. W05404.

D. Mao, L. Wan, T.-C. J. Yeh, C.-H. Lee, K.-C. Hsu, J.-C. Wen, and W. Lu. (2011). A revisit of drawdown behavior during pumping in unconfined aquifers. *Water Resources Research*, 47(5):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2010WR009326. W05502.

D. Mao, T. C. J. Yeh, L. Wan, K.-C. Hsu, C.-H. Lee, and J.-C. Wen. (2013a). Necessary conditions for inverse modeling of flow through variably saturated

porous media. *Advances in Water Resources*, 52:50–61. ISSN 03091708. doi: 10.1016/j.advwatres.2012.08.001. URL `http://www.sciencedirect.com/science/article/pii/S0309170812002217`.

D. Mao, T. C. J. Yeh, L. Wan, C. H. Lee, K. C. Hsu, J. C. Wen, and W. Lu. (2013b). Cross-correlation analysis and information content of observed heads during pumping in unconfined aquifers. *Water Resources Research*, 49(2):713–731. ISSN 00431397. doi: 10.1002/wrcr.20066.

D. Mao, T.-C. J. Yeh, L. Wan, J.-C. Wen, W. Lu, C.-H. Lee, and K.-C. Hsu. (2013c). Joint interpretation of sequential pumping tests in unconfined aquifers. *Water Resources Research*, 49(4):1782–1796. ISSN 1944-7973. doi: 10.1002/wrcr.20129.

Y. Mualem. (jun 1976). A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, 12(3):513–522. ISSN 1944-7973. doi: 10.1029/WR012i003p00513. URL `http://dx.doi.org/10.1029/WR012i003p00513`.

S. P. Neuman. (1972). Theory of flow in unconfined aquifers considering delayed response of the water table. *Water Resources Research*, 8(4):1031–1045. ISSN 1944-7973. doi: 10.1029/WR008i004p01031.

S. P. Neuman. (1974). Effect of partial penetration on flow in unconfined aquifers considering delayed gravity response. *Water Resources Research*, 10(2):303–312. ISSN 1944-7973. doi: 10.1029/WR010i002p00303.

C.-F. Ni, T.-C. J. Yeh, and J.-S. Chen. (2009). Cost-effective hydraulic tomography surveys for predicting flow and transport in heterogeneous aquifers. *Environmental Science & Technology*, 43(10):3720–3727. doi: 10.1021/es8024098.

G. I. Nwankwor, J. A. Cherry, and R. W. Gillham. (1984). A comparative study of specific yield determinations for a shallow sand aquifer. *Ground Water*, 22(6):764–772. ISSN 1745-6584. doi: 10.1111/j.1745-6584.1984.tb01445.x.

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

R. Rojas, L. Feyen, and A. Dassargues. (2008). Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research*, 44(12):1–16. ISSN 00431397. doi: 10.1029/2008WR006908.

R. Rojas, L. Feyen, O. Batelaan, and A. Dassargues. (2010). On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling. *Water Resources Research*, 46(8):1–20. ISSN 00431397. doi: 10.1029/2009WR008822.

D. Russo. (1988). Determining soil hydraulic properties by parameter estimation: On the selection of a model for the hydraulic properties. *Water Resources Research*, 24(3): 453–459. ISSN 1944-7973. doi: 10.1029/WR024i003p00453. URL `http://dx.doi.org/10.1029/WR024i003p00453`.

A. Schöniger, W. Nowak, and H. J. Hendricks Franssen. (2012). Parameter estimation by ensemble Kalman filters with transformed data: Approach and application to hydraulic tomography. *Water Resources Research*, 48(4):1–18. ISSN 00431397. doi: 10.1029/2011WR010462.

A. Schöniger, W. A. Illman, T. Wöhling, and W. Nowak. (2015). Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. *Journal of Hydrology*, 531:96–110. ISSN 00221694. doi: 10.1016/j.jhydrol.2015.07.047. URL `http://linkinghub.elsevier.com/retrieve/pii/S0022169415005533`.

L. A. Sebol. (2000). *Determination of groundwater age using CFC's in three shallow aquifers in Southern Ontario*. PhD thesis, University of Waterloo, Waterloo, Ontario, Canada.

S. Straface, T. C. J. Yeh, J. Zhu, S. Troisi, and C. H. Lee. (2007). Sequential aquifer tests at a well field, Montalto Uffugo Scalo, Italy. *Water Resources Research*, 43(7): 1–13. ISSN 00431397. doi: 10.1029/2006WR005287.

S. Strebelle. (2002). Conditional somulation of complex geological structures using multiple-point geostatistics. *Math. Geol.*, 34(1):1–22. doi: 10.1023/A:1014009426274.

E. A. Sudicky. (dec 1986). A natural gradient experiment on solute transport in a sand aquifer: Spatial variability of hydraulic conductivity and its role in the dispersion process. *Water Resources Research*, 22(13):2069–2082. ISSN 1944-7973. doi: 10.1029/WR022i013p02069. URL `http://dx.doi.org/10.1029/WR022i013p02069`.

E. a. Sudicky, W. a. Illman, I. K. Goltz, J. J. Adams, and R. G. Mclaren. (2010). Heterogeneity in hydraulic conductivity and its role on the macroscale transport of a solute plume: From measurements to a practical application of stochastic flow and transport theory. *Water Resources Research*, 46(1):W01508 (1–16). ISSN <null>. doi: 10.1029/2008WR007558. URL `papers2://publication/doi/10.1029/2008WR007558`.

R. Sun, T. C. J. Yeh, D. Mao, M. Jin, W. Lu, and Y. Hao. (2013). A temporal sampling strategy for hydraulic tomography analysis. *Water Resources Research*, 49(7):3881–3896. ISSN 00431397. doi: 10.1002/wrcr.20337.

V. Teles, F. Delay, and G. De Marsily. (2004). Comparison of genesis and geostatistical methods for characterizing the heterogeneity of alluvial media: Groundwater flow and transport simulations. *Journal of Hydrology*, 294(1-3):103–121. ISSN 00221694. doi: 10.1016/j.jhydrol.2003.11.041.

C. V. Theis. (1935). The relation between the lowering of the piezometric surface and the rate and duration of discharge of a well using ground-water storage. *Eos, Transactions American Geophysical Union*, 16(2):519–524. ISSN 2324-9250. doi: 10.1029/TR016i002p00519.

L. Troldborg, J. C. Refsgaard, K. H. Jensen, and P. Engesgaard. (2007). The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. *Hydrogeology Journal*, 15(5):843–860. ISSN 14312174. doi: 10.1007/s10040-007-0192-y.

M. van Genuchten. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci Soc Am J*, (44):892–898.

G. Weissmann, S. Carle, and G. Fogg. (1999). Three-dimensional hydrofacies modeling based on soil surveys and transition probability. *Water Resources Research*, 35(6):1761–1770. ISSN 0043-1397. doi: 10.1029/1999WR900048.

J. C. Wen, C. M. Wu, T. C. J. Yeh, and C. M. Tseng. (2010). Estimation of effective aquifer hydraulic properties from an aquifer test with multi-well observations (Taiwan). *Hydrogeology Journal*, 18(5):1143–1155. ISSN 14312174. doi: 10.1007/s10040-010-0577-1.

C. M. Wu, T. C. J. Yeh, J. Zhu, H. L. Tim, N. S. Hsu, C. H. Chen, and A. F. Sancho. (2005). Traditional analysis of aquifer tests: Comparing apples to oranges? *Water Resources Research*, 41(9):1–12. ISSN 00431397. doi: 10.1029/2004WR003717.

J. Xiang, T.-C. J. Yeh, C.-H. Lee, K.-C. Hsu, and J.-C. Wen. (2009). A simultaneous successive linear estimator and a guide for hydraulic tomography analysis. *Water Resources Research*, 45(2):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2008WR007180. W02432.

T.-C. J. Yeh and S. Liu. (2000). Hydraulic tomography: Development of a new aquifer test method. *Water Resources Research*, 36(8):2095–2105. ISSN 1944-7973. doi: 10.1029/2000WR900114.

T.-C. J. Yeh, R. Srivastava, A. Guzman, and T. Harter. (1993). A numerical model for water flow and chemical transport in variably saturated porous media. *Ground Water*, 31(4):634–644. ISSN 1745-6584. doi: 10.1111/j.1745-6584.1993.tb00597.x. URL `http://dx.doi.org/10.1111/j.1745-6584.1993.tb00597.x`.

T. C. J. Yeh, M. Jin, and S. Hanna. (1996). An Iterative Stochastic Inverse Method: Conditional Effective Transmissivity and Hydraulic Head Fields. *Water Resources Research*, 32(1):85–92. ISSN 00431397. doi: 10.1029/95WR02869. URL `http://doi.wiley.com/10.1029/95WR02869`.

T.-C. J. Yeh, J. Xiang, R. M. Suribhatla, K.-C. Hsu, C.-H. Lee, and J.-C. Wen. (may 2009). River stage tomography: A new approach for characterizing groundwater basins. *Water Resources Research*, 45(5):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2008WR007233. URL `http://dx.doi.org/10.1029/2008WR007233`.

T.-C. J. Yeh, D. Mao, L. Wan, C.-H. Lee, J.-C. Wen, and W. Lu. (2012). Replies to comments on "a revisit of drawdown behavior during pumping in unconfined aquifers" by neuman and mishra. *Water Resources Research*, 48(2):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2011WR011153. W02802.

D. Yin and W. A. Illman. (2009). Hydraulic tomography using temporal moments of drawdown recovery data: A laboratory sandbox study. *Water Resources Research*, 45 (1):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2007WR006623. W01502.

Y. Zha, T. C. J. Yeh, W. A. Illman, T. Tanaka, P. Bruines, H. Onoe, and H. Saegusa. (2015). What does hydraulic tomography tell us about fractured geological media? A field study and synthetic experiments. *Journal of Hydrology*, 531:17–30. ISSN 00221694. doi: 10.1016/j.jhydrol.2015.06.013. URL `http://dx.doi.org/10.1016/j.jhydrol.2015.06.013`.

Y. Zha, T.-C. J. Yeh, W. A. Illman, T. Tanaka, P. Bruines, H. Onoe, H. Saegusa, D. Mao, S. Takeuchi, and J.-C. Wen. (apr 2016). An Application of Hydraulic Tomography to a Large-Scale Fractured Granite Site, Mizunami, Japan. *Groundwater*, pages n/a–n/a. ISSN 1745-6584. doi: 10.1111/gwat.12421. URL `http://dx.doi.org/10.1111/gwat.12421`.

Z. Zhao, W. A. Illman, T.-C. J. Yeh, S. J. Berg, and D. Mao. (2015). Validation of hydraulic tomography in an unconfined aquifer: A controlled sandbox study. *Water Resources Research*, 51(6):4137–4155. ISSN 1944-7973. doi: 10.1002/2015WR016910. URL `http://dx.doi.org/10.1002/2015WR016910`.

Z. Zhao, W. A. Illman, and S. J. Berg. (2016). On the importance of geological data for hydraulic tomography analysis: Laboratory sandbox study. *Journal of Hydrology*. ISSN 0022-1694. doi: http://dx.doi.org/10.1016/j.jhydrol.2016.08.061. URL `http://www.sciencedirect.com/science/article/pii/S0022169416305510`.

Y. Zhou, D. Lim, F. Cupola, and M. Cardiff. (2016). Aquifer imaging with pressure waves-Evaluation of low-impact characterization through sandbox experiments. *Water Resources Research*, pages n/a–n/a. ISSN 00431397. doi: 10.1002/2015WR017751. URL `http://doi.wiley.com/10.1002/2015WR017751`.

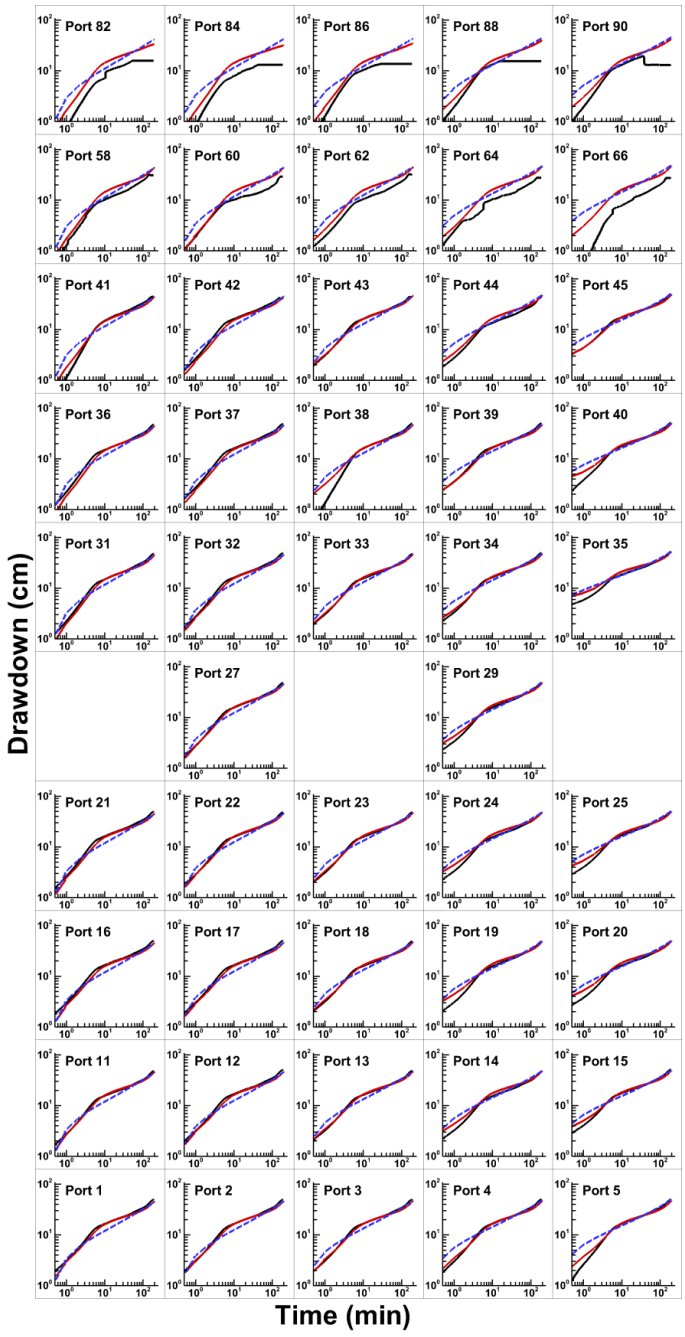
J. Zhu and T.-C. J. Yeh. (2005). Characterization of aquifer heterogeneity using transient hydraulic tomography. *Water Resources Research*, 41(7):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2004WR003790. W07028.

J. Zhu and T.-C. J. Yeh. (2006). Analysis of hydraulic tomography using temporal moments of drawdown recovery data. *Water Resources Research*, 42(2):n/a–n/a. ISSN 1944-7973. doi: 10.1029/2005WR004309. W02403.

J. Zhu, T.-C. J. Yeh, and D. Mao. (2011). Hydraulic tomography to characterize heterogeneity of unconfined aquifers. *J. Nanjing Univ.*, 47(3):252–264.
doi:10.1016/j.jhydrol.2016.08.061

Supplementary Information A

a)


Port 82
Port 84
Port 86
Port 88
Port 90
Port 58
Port 60
Port 62
Port 64
Port 66
Port 41
Port 42
Port 43
Port 44
Port 45
Port 36
Port 37
Port 38
Port 39
Port 40
Port 31
Port 32
Port 33
Port 34
Port 35
Port 27
Port 29
Port 21
Port 22
Port 23
Port 24
Port 25
Port 16
Port 17
Port 18
Port 19
Port 20
Port 11
Port 12
Port 13
Port 14
Port 15
Port 1
Port 2
Port 3
Port 4
Port 5
Drawdown (cm)
Time (min)

b)

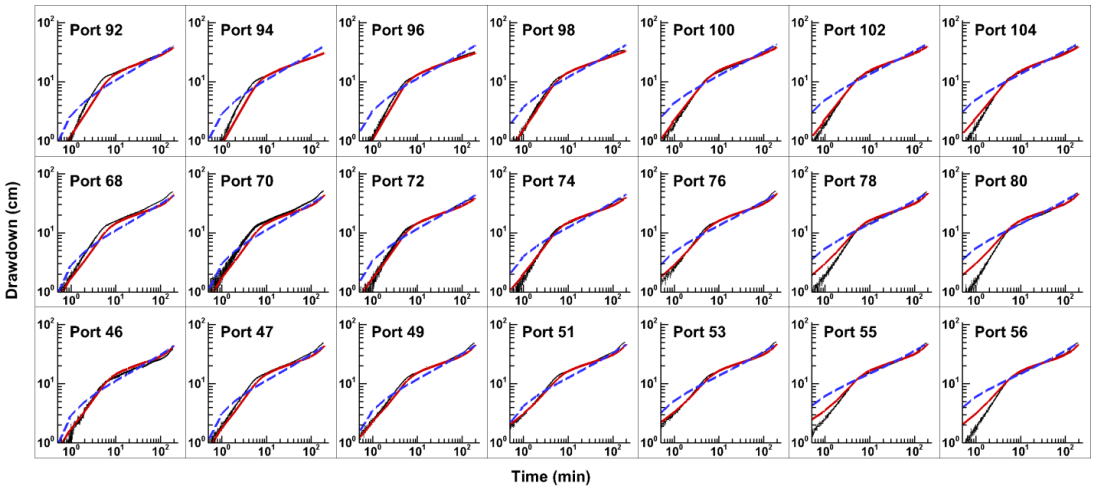
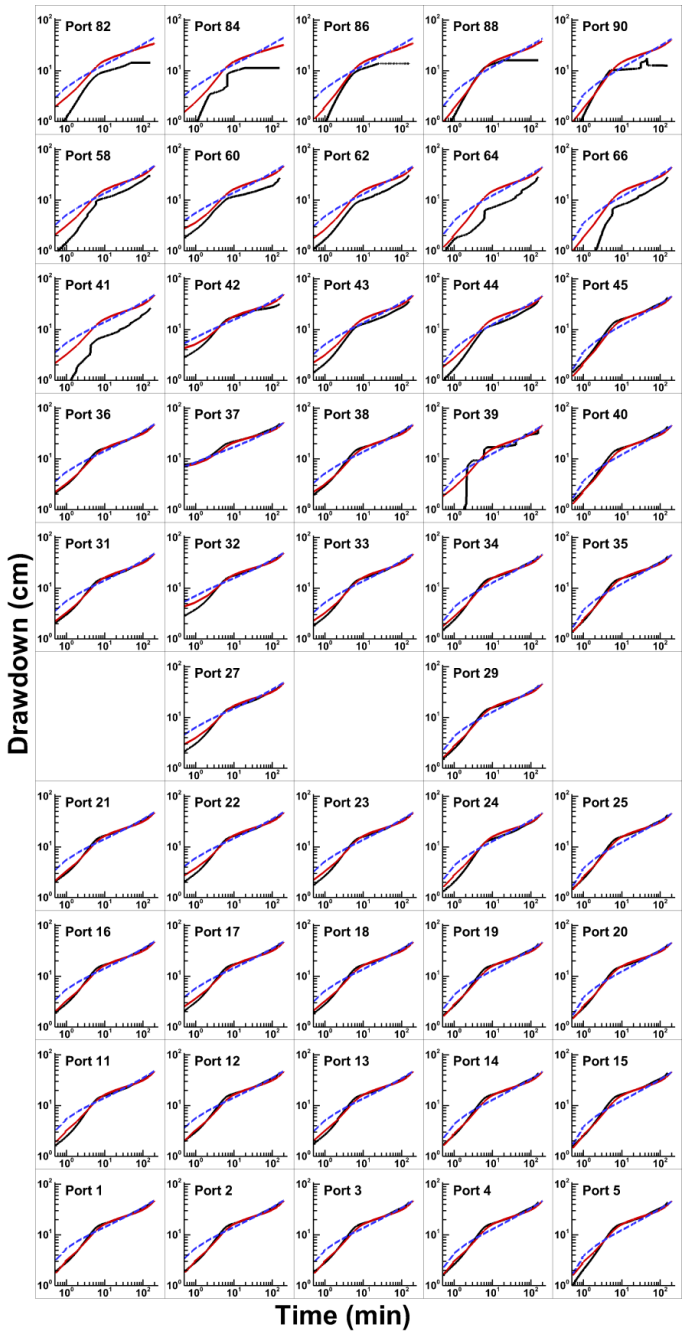


Figure A1. Simulated and observed drawdown at: a) pressure transducer and b) tensiometer ports during a pumping test at port 35. Black solid lines are observed data, red lines are simulated drawdown using results from HT; and blue dashed lines are simulated drawdown using the homogeneous assumption. The layout of all the ports in this figure is identical to the layout on the sandbox shown in Figure 1.

a)


Port 82
Port 84
Port 86
Port 88
Port 90
Port 58
Port 60
Port 62
Port 64
Port 66
Port 41
Port 42
Port 43
Port 44
Port 45
Port 36
Port 37
Port 38
Port 39
Port 40
Port 31
Port 32
Port 33
Port 34
Port 35
Port 27
Port 29
Port 21
Port 22
Port 23
Port 24
Port 25
Port 16
Port 17
Port 18
Port 19
Port 20
Port 11
Port 12
Port 13
Port 14
Port 15
Port 1
Port 2
Port 3
Port 4
Port 5
Drawdown (cm)
Time (min)

b)

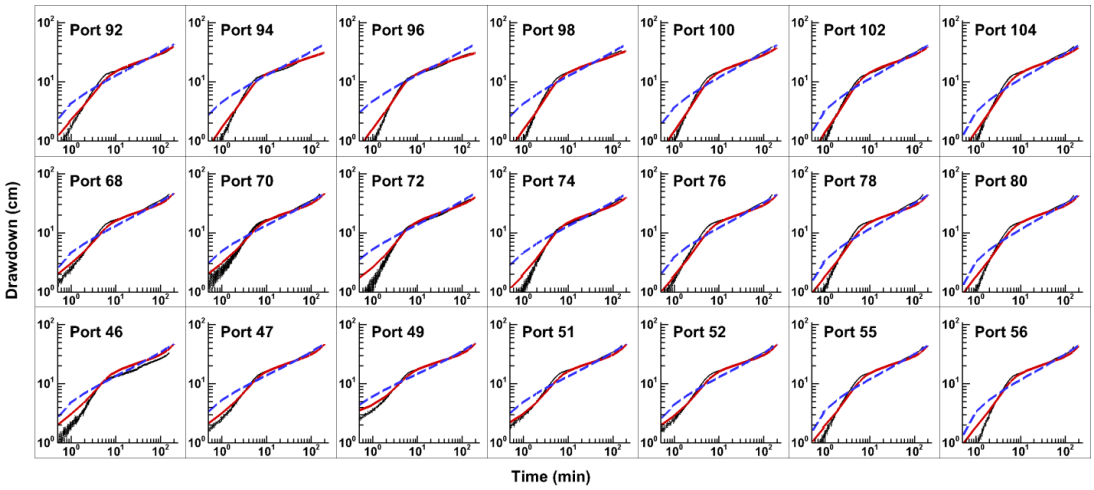
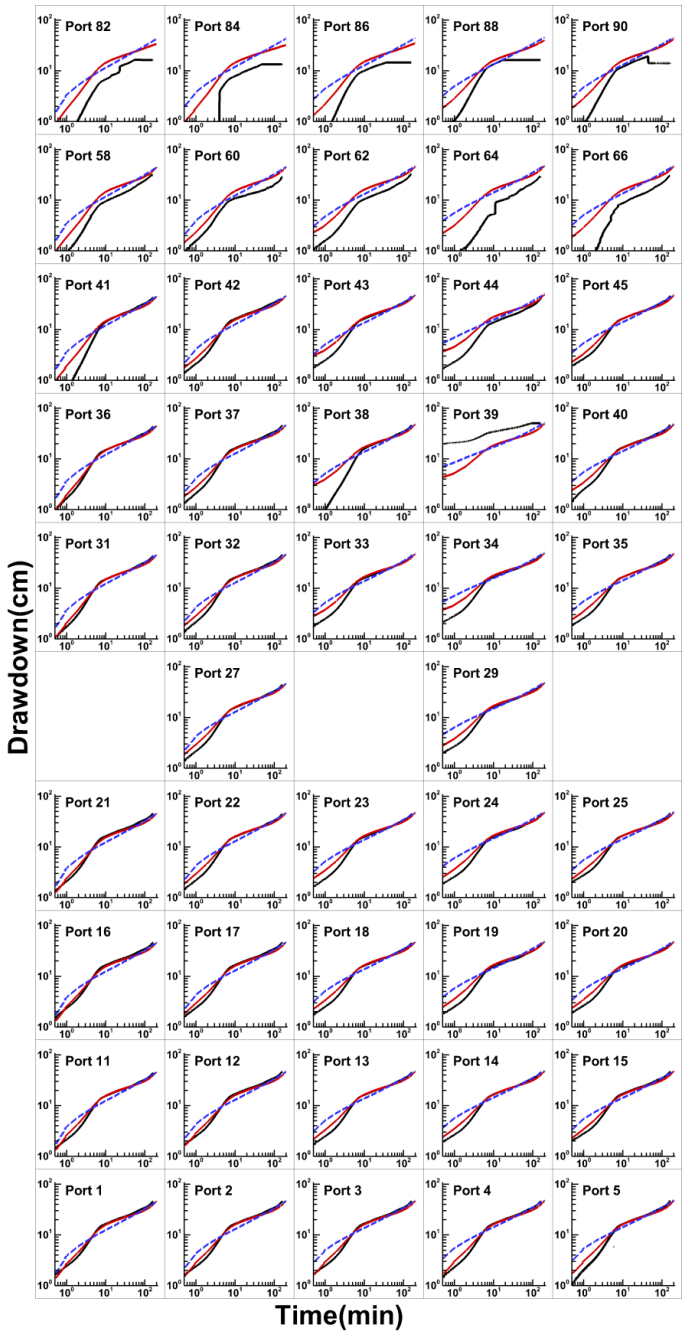


Figure A2. Simulated and observed drawdown at: a) pressure transducer and b) tensiometer ports during a pumping test at port 37. Black solid lines are observed data, red lines are simulated drawdown using results from HT; and blue dashed lines are simulated drawdown using the homogeneous assumption. The layout of all the ports in this figure is identical to the layout on the sandbox shown in Figure 1.

a)


Port 82
Port 84
Port 86
Port 88
Port 90
Port 58
Port 60
Port 62
Port 64
Port 66
Port 41
Port 42
Port 43
Port 44
Port 45
Port 36
Port 37
Port 38
Port 39
Port 40
Port 31
Port 32
Port 33
Port 34
Port 35
Port 27
Port 29
Port 21
Port 22
Port 23
Port 24
Port 25
Port 16
Port 17
Port 18
Port 19
Port 20
Port 11
Port 12
Port 13
Port 14
Port 15
Port 1
Port 2
Port 3
Port 4
Port 5
Drawdown(cm)
Time(min)

b)

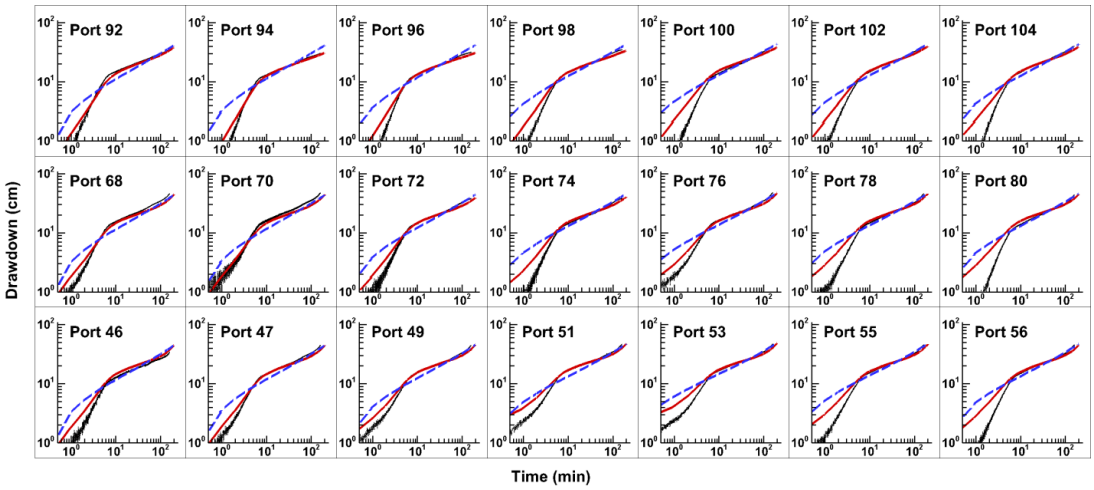


Figure A3. Simulated and observed drawdown at: a) pressure transducer and b) tensiometer ports during a pumping test at port 39. Black solid lines are observed data, red lines are simulated drawdown using results from HT; and blue dashed lines are simulated drawdown using the homogeneous assumption. The layout of all the ports in this figure is identical to the layout on the sandbox shown in Figure 1.

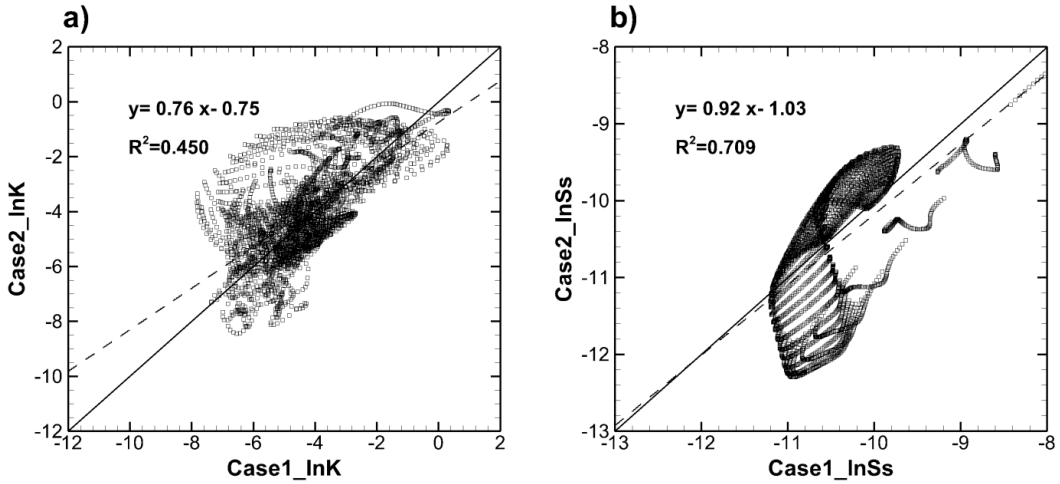


Figure A4. Scatterplots of estimated tomograms between Case 2 and Case1, a) ln K, b) ln S_s, c) ln α, d) ln θ_s. The solid line is a 1:1 line indicating a perfect match. The dash line is a best fit line. The linear fit results are also included on each plot.

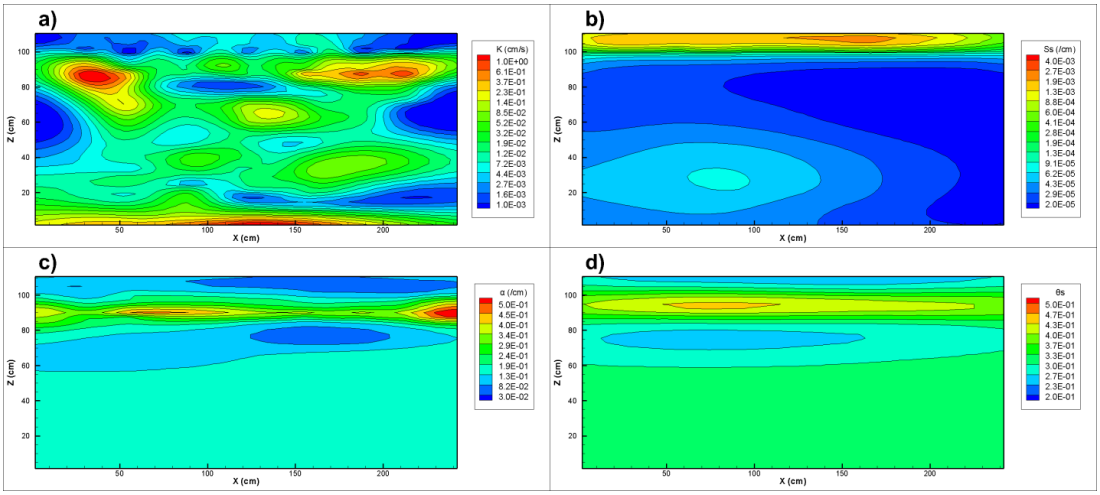


Figure A5. Estimated tomograms of: a) K, b) S_s, c) α, d) θ_s using different initial values (case 3). Spatial variances of the estimated tomograms are $\sigma^2_{\ln K}$= 2.59, $\sigma^2_{\ln Ss}$= 1.29, $\sigma^2_{\ln \alpha}$ = 0.12, $\sigma^2_{\ln \theta s}$ = 0.01.

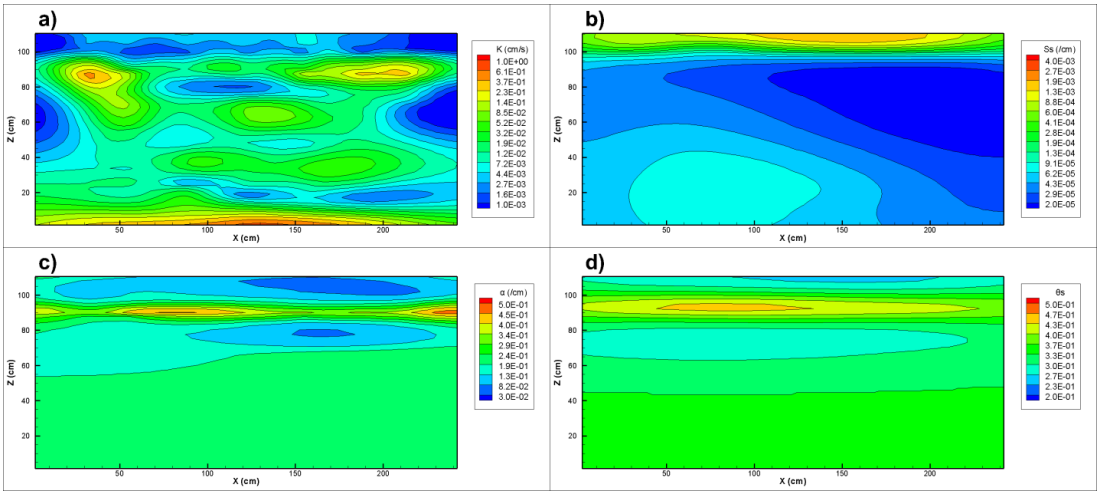


Figure A6. Estimated tomograms of: a) K, b) S_s, c) α, d) θ_s using different initial values (case 4). Spatial variances of the estimated tomograms are $\sigma^2_{\ln K}$= 1.97, $\sigma^2_{\ln Ss}$= 1.05, $\sigma^2_{\ln \alpha}$ = 0.10, $\sigma^2_{\ln \theta s}$ = 0.01.

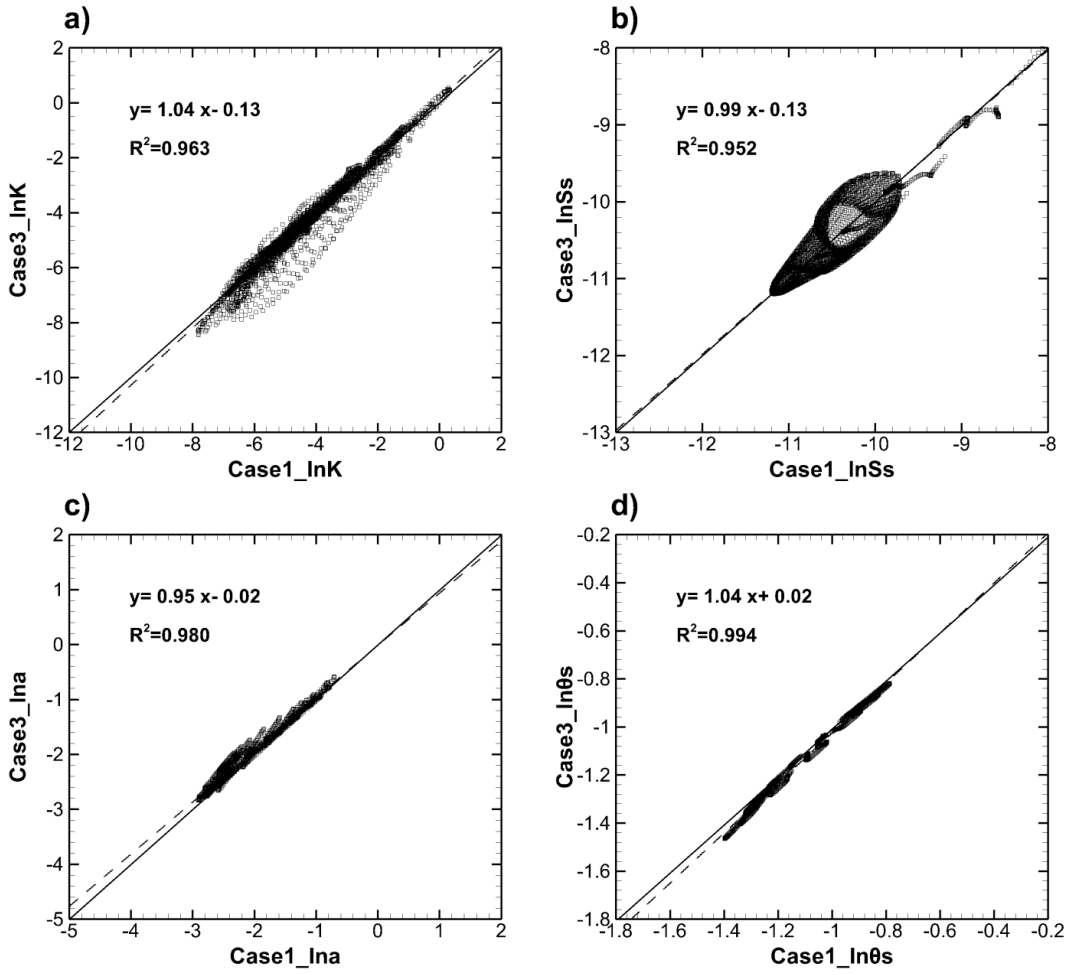


Figure A7. Scatterplots of estimated tomograms between Case 3 and Case 1: a) ln K, b) ln S_s, c) ln α, d) ln θ_s. The solid line is a 1:1 line indicating a perfect match. The dash line is a best fit line. The linear fit results are also included on each plot.

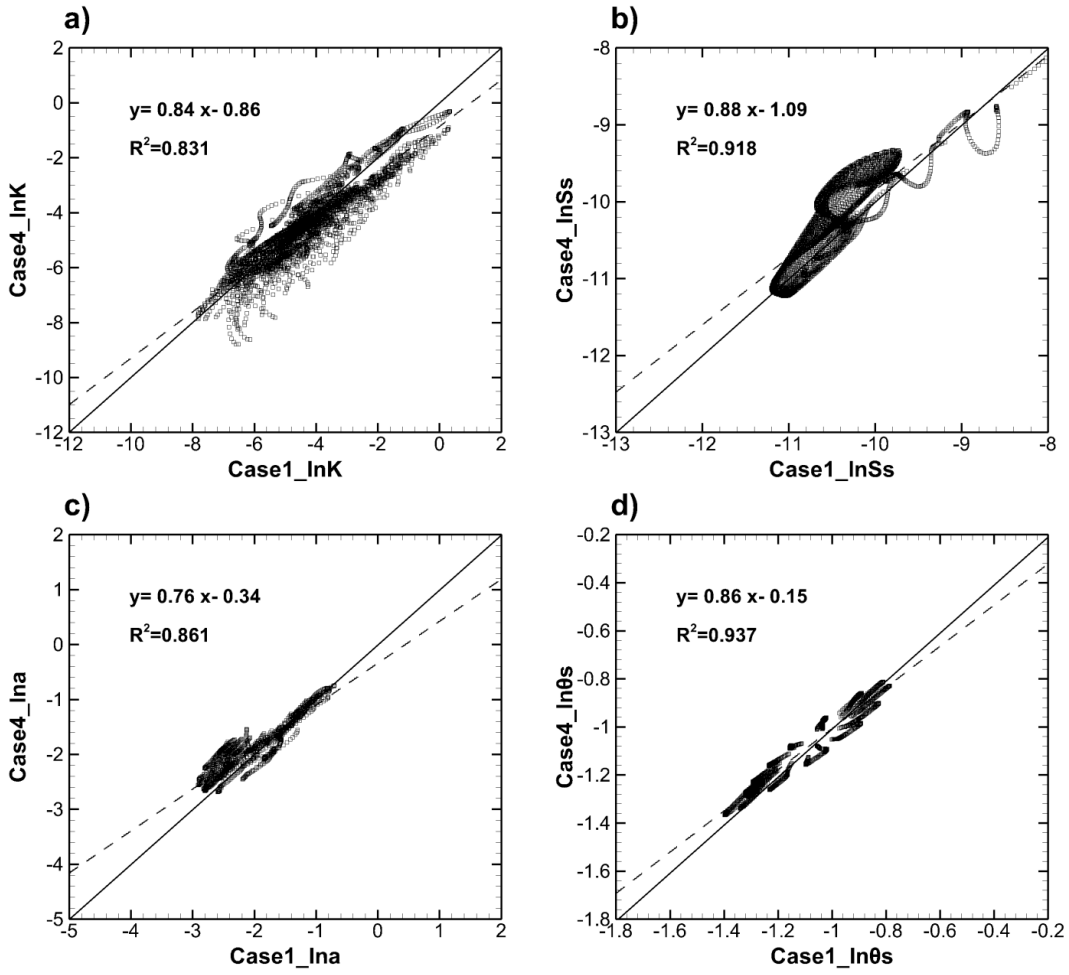


Figure A8. Scatterplots of estimated tomograms between Case 4 and Case1, a) ln K, b) ln S_s, c) ln α, d) ln θ_s. The solid line is a 1:1 line indicating a perfect match. The dash line is a best fit line. The linear fit results are also included on each plot.

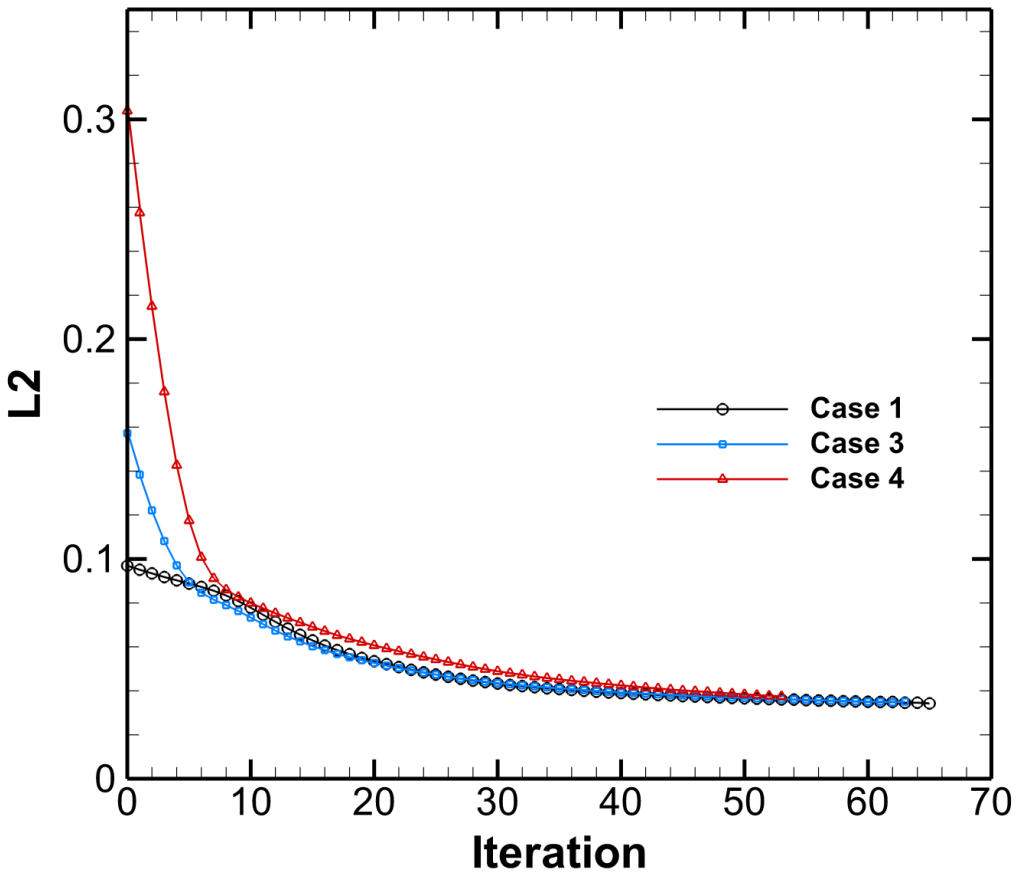


Figure A9. *L2* norms of the head as a function of iteration for: Case1 in black circles; Case 3 in blue squares; Case 4 in red triangles.

Supplementary Information B

Table B1. Sand type and K (cm/s) estimates for each layer in the sandbox aquifer*.

Layer	Sand type	d_{50} (mm)	Core permeameter
1	20/30	0.750	3.2×10^{-2}
2	4030	0.355	5.3×10^{-2}
3	F-85	0.151	7.1×10^{-2}
4	20/40	0.578	5.7×10^{-2}
5	Mix	N/A	6.2×10^{-2} **
6	Mix	N/A	8.2×10^{-2}
7	#12	0.525	1.3×10^{-1}
8	F32	0.504	1.3×10^{-1}
9	20/40	0.578	8.7×10^{-2}
10	F-65	0.204	1.1×10^{-1}
11	#12	0.525	1.4×10^{-1}
12	16/30	0.872	3.4×10^{-2}
13	20/30	0.750	2.6×10^{-1}
14	F-75	0.174	9.8×10^{-2}
15	20/40	0.578	8.6×10^{-2}
16	Mix	N/A	4.2×10^{-2}
17	F-85	0.151	4.5×10^{-2}
18	20/30	0.750	1.5×10^{-1}

Note: The layers labelled "mix" consisted of equal volumes of #14, F75, and 16/30 sands.
N/A denotes that data are not available for using mixed sand types.
* These core permeameter test results are adopted from Illman et al., (2010).
** This K of layer 5 is estimated as the mean value of K of layer 6 and 16 with same sand type.

Table B2. The *K* (cm/s) values and corresponding 95% confidence interval limits estimated for each geological model in Case 1.

Layer	Geo-Good			Geo-Poor1			Geo-Poor2			Geo-Poor3		
	Estimate	95% confidence Interval		Estimate	95% confidence Interval		Estimate	95% confidence Interval		Estimate	95% confidence Interval	
		Lower limit	Upper limit		Lower limit	Upper limit		Lower limit	Upper limit		Lower limit	Upper limit
1	1.33×10^{-01}	9.47×10^{-02}	1.87×10^{-01}	$1.32\times10^{+00}$	9.75×10^{-01}	$1.80\times10^{+00}$	1.69×10^{-01}	1.47×10^{-01}	1.96×10^{-01}	3.18×10^{-01}	2.06×10^{-01}	4.89×10^{-01}
2	5.18×10^{-02}	2.55×10^{-02}	1.05×10^{-01}	5.30×10^{-02}	4.04×10^{-02}	6.95×10^{-02}				3.81×10^{-02}	9.71×10^{-03}	1.50×10^{-01}
3	1.00×10^{-04}	6.27×10^{-11}	$1.60\times10^{+02}$	1.16×10^{-01}	9.56×10^{-02}	1.42×10^{-01}				9.62×10^{-02}	7.81×10^{-02}	1.19×10^{-01}
4	2.44×10^{-01}	1.94×10^{-01}	3.08×10^{-01}	3.32×10^{-02}	2.00×10^{-02}	5.51×10^{-02}				1.96×10^{-01}	9.93×10^{-02}	3.87×10^{-01}
5	3.89×10^{-02}	2.93×10^{-02}	5.16×10^{-02}	3.36×10^{-02}	2.37×10^{-02}	4.77×10^{-02}	5.94×10^{-02}	5.11×10^{-02}	6.91×10^{-02}	3.62×10^{-02}	2.19×10^{-02}	5.98×10^{-02}
6	8.77×10^{-03}	5.98×10^{-03}	1.29×10^{-02}	2.60×10^{-04}	2.26×10^{-07}	2.98×10^{-01}				2.94×10^{-02}	2.01×10^{-02}	4.30×10^{-02}
7	1.25×10^{-01}	8.51×10^{-02}	1.82×10^{-01}	3.04×10^{-02}	2.12×10^{-02}	4.35×10^{-02}				2.70×10^{-02}	1.60×10^{-02}	4.55×10^{-02}
8	3.51×10^{-02}	2.82×10^{-02}	4.37×10^{-02}	4.18×10^{-02}	2.71×10^{-02}	6.45×10^{-02}				1.01×10^{-01}	3.09×10^{-02}	3.29×10^{-01}
9	1.66×10^{-01}	1.30×10^{-01}	2.12×10^{-01}	2.66×10^{-01}	2.11×10^{-01}	3.36×10^{-01}				2.49×10^{-01}	1.81×10^{-01}	3.44×10^{-01}
10	1.85×10^{-02}	1.65×10^{-02}	2.06×10^{-02}	3.93×10^{-02}	3.44×10^{-02}	4.49×10^{-02}	3.41×10^{-02}	2.80×10^{-02}	4.15×10^{-02}	1.64×10^{-01}	1.17×10^{-01}	2.31×10^{-01}
11	1.15×10^{-01}	4.90×10^{-02}	2.70×10^{-01}	1.14×10^{-01}	6.81×10^{-02}	1.91×10^{-01}				4.57×10^{-02}	3.44×10^{-02}	6.07×10^{-02}
12	9.18×10^{-02}	4.81×10^{-02}	1.75×10^{-01}	4.92×10^{-02}	2.90×10^{-02}	8.36×10^{-02}				2.18×10^{-02}	1.72×10^{-02}	2.76×10^{-02}
13	4.45×10^{-01}	3.78×10^{-01}	5.25×10^{-01}	5.22×10^{-01}	4.13×10^{-01}	6.59×10^{-01}				3.91×10^{-01}	2.55×10^{-01}	5.98×10^{-01}
14	2.82×10^{-02}	2.36×10^{-02}	3.37×10^{-02}	5.38×10^{-03}	2.77×10^{-03}	1.05×10^{-02}	6.66×10^{-02}	5.90×10^{-02}	7.52×10^{-02}	8.85×10^{-02}	5.57×10^{-02}	1.41×10^{-01}
15	1.07×10^{-01}	8.04×10^{-02}	1.42×10^{-01}	9.87×10^{-02}	7.33×10^{-02}	1.33×10^{-01}				7.55×10^{-01}	4.95×10^{-01}	$1.15\times10^{+00}$
16	3.88×10^{-02}	3.28×10^{-02}	4.59×10^{-02}	2.68×10^{-02}	2.30×10^{-02}	3.13×10^{-02}				1.59×10^{-02}	1.36×10^{-02}	1.85×10^{-02}
17	3.70×10^{-02}	3.13×10^{-02}	4.39×10^{-02}	1.95×10^{-02}	1.44×10^{-02}	2.66×10^{-02}				7.46×10^{-01}	2.61×10^{-02}	$2.14\times10^{+01}$
18	3.60×10^{-01}	3.04×10^{-01}	4.25×10^{-01}	2.67×10^{-01}	2.32×10^{-01}	3.06×10^{-01}	2.29×10^{-01}	1.89×10^{-01}	2.78×10^{-01}	3.29×10^{-01}	2.25×10^{-01}	4.82×10^{-01}

Table B3. The *K* (cm/s) values and corresponding 95% confidence interval limits estimated for each geological model in Case 2.

Layer	Geo-Good			Geo-Poor1			Geo-Poor2			Geo-Poor3		
	Estimate	95% confidence Interval		Estimate	95% confidence Interval		Estimate	95% confidence Interval		Estimate	95% confidence Interval	
		Lower limit	Upper limit		Lower limit	Upper limit		Lower limit	Upper limit		Lower limit	Upper limit
1	2.56×10^{-02}	2.18×10^{-05}	$3.01\times10^{+01}$	$4.11\times10^{+00}$	3.24×10^{-02}	$5.22\times10^{+02}$	2.34×10^{-01}	1.76×10^{-01}	3.10×10^{-01}	8.95×10^{-01}	5.52×10^{-01}	$1.45\times10^{+00}$
2	9.87×10^{-04}	1.81×10^{-10}	$5.37\times10^{+03}$	3.79×10^{-02}	1.53×10^{-02}	9.36×10^{-02}				4.48×10^{-02}	3.32×10^{-03}	6.06×10^{-01}
3	1.00×10^{-04}	2.28×10^{-08}	4.38×10^{-01}	9.91×10^{-02}	1.64×10^{-02}	5.99×10^{-01}				1.44×10^{-01}	1.02×10^{-01}	2.04×10^{-01}
4	3.35×10^{-01}	2.69×10^{-01}	4.18×10^{-01}	1.19×10^{-01}	5.40×10^{-03}	$2.63\times10^{+00}$				7.53×10^{-02}	1.94×10^{-02}	2.92×10^{-01}
5	3.61×10^{-02}	2.48×10^{-02}	5.24×10^{-02}	4.93×10^{-02}	1.29×10^{-02}	1.88×10^{-01}	7.63×10^{-02}	5.24×10^{-02}	1.11×10^{-01}	5.14×10^{-02}	1.64×10^{-02}	1.61×10^{-01}
6	1.62×10^{-02}	8.39×10^{-03}	3.11×10^{-02}	1.96×10^{-02}	1.29×10^{-11}	$3.00\times10^{+07}$				3.37×10^{-02}	1.65×10^{-02}	6.88×10^{-02}
7	7.51×10^{-02}	4.28×10^{-02}	1.32×10^{-01}	2.10×10^{-02}	8.03×10^{-03}	5.50×10^{-02}				1.32×10^{-01}	4.94×10^{-04}	$3.51\times10^{+01}$
8	3.04×10^{-02}	2.56×10^{-02}	3.62×10^{-02}	3.35×10^{-02}	1.54×10^{-02}	7.32×10^{-02}				3.76×10^{-01}	3.23×10^{-05}	$4.38\times10^{+03}$
9	1.94×10^{-01}	1.46×10^{-01}	2.60×10^{-01}	1.89×10^{-01}	1.08×10^{-01}	3.31×10^{-01}				2.02×10^{-01}	4.35×10^{-02}	9.38×10^{-01}
10	1.98×10^{-02}	1.65×10^{-02}	2.36×10^{-02}	4.05×10^{-02}	3.11×10^{-02}	5.27×10^{-02}	1.91×10^{-02}	1.15×10^{-02}	3.16×10^{-02}	2.89×10^{-01}	6.60×10^{-02}	$1.26\times10^{+00}$
11	3.76×10^{-01}	1.65×10^{-01}	8.60×10^{-01}	1.77×10^{-01}	5.77×10^{-02}	5.42×10^{-01}				4.04×10^{-02}	3.55×10^{-03}	4.60×10^{-01}
12	1.32×10^{-01}	3.10×10^{-02}	5.62×10^{-01}	4.40×10^{-02}	8.52×10^{-03}	2.27×10^{-01}				1.29×10^{-02}	2.10×10^{-03}	7.88×10^{-02}
13	4.83×10^{-01}	3.87×10^{-01}	6.04×10^{-01}	$1.93\times10^{+00}$	9.53×10^{-01}	$3.90\times10^{+00}$				8.93×10^{-02}	4.14×10^{-06}	$1.93\times10^{+03}$
14	2.07×10^{-02}	9.13×10^{-03}	4.68×10^{-02}	3.72×10^{-03}	3.69×10^{-04}	3.74×10^{-02}	8.66×10^{-02}	6.14×10^{-02}	1.22×10^{-01}	5.79×10^{-02}	5.68×10^{-03}	5.90×10^{-01}
15	8.08×10^{-02}	6.26×10^{-02}	1.04×10^{-01}	1.40×10^{-01}	6.68×10^{-02}	2.94×10^{-01}				1.83×10^{-02}	1.56×10^{-03}	2.15×10^{-01}
16	2.28×10^{-02}	9.74×10^{-03}	5.33×10^{-02}	1.61×10^{-02}	8.83×10^{-03}	2.93×10^{-02}				5.64×10^{-02}	1.10×10^{-03}	$2.88\times10^{+00}$
17	4.46×10^{-02}	3.47×10^{-02}	5.73×10^{-02}	1.26×10^{-02}	3.18×10^{-03}	4.95×10^{-02}				7.74×10^{-01}	2.64×10^{-05}	$2.27\times10^{+04}$
18	3.46×10^{-01}	2.09×10^{-01}	5.73×10^{-01}	2.81×10^{-01}	1.42×10^{-01}	5.54×10^{-01}	1.40×10^{-01}	7.06×10^{-02}	2.79×10^{-01}	8.44×10^{-02}	1.58×10^{-02}	4.50×10^{-01}

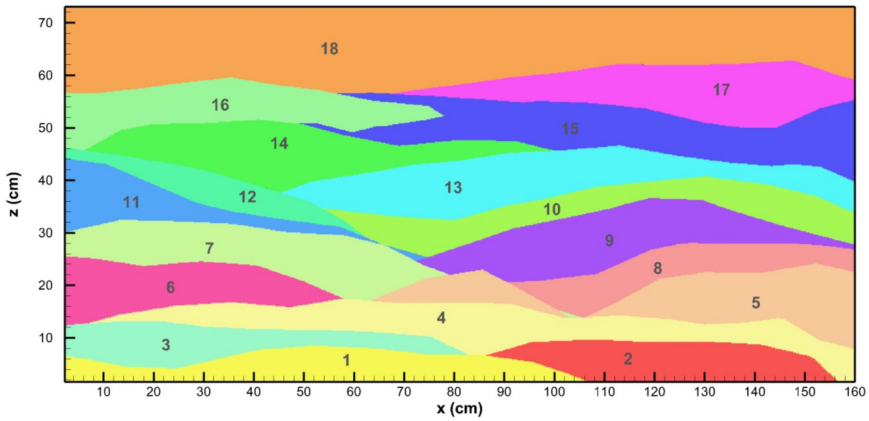


Fig. B1. "Perfect" geological model built based on accurate stratigraphy shown in Fig. 1a.

L1	Port 2	Port 5	Port 14	Port 17	Port 26	Port 29	Port 44	Port 47	Average	Rank
Geo_good	0.022	0.067	0.074	0.094	0.116	0.106	0.121	0.113	0.089	2
Geo_poor1	0.036	0.080	0.086	0.121	0.178	0.130	0.128	0.116	0.109	3
Geo_poor2	0.077	0.120	0.146	0.260	0.262	0.233	0.340	0.178	0.202	5
Geo_poor3	0.033	0.066	0.114	0.128	0.206	0.158	0.241	0.166	0.139	4
SimSLE	0.019	0.021	0.033	0.033	0.058	0.033	0.024	0.024	0.031	1

L2	Port 2	Port 5	Port 14	Port 17	Port 26	Port 29	Port 44	Port 47	Average	Rank
Geo_good	0.001	0.006	0.025	0.020	0.026	0.020	0.030	0.024	0.019	2
Geo_poor1	0.002	0.009	0.015	0.044	0.050	0.048	0.031	0.026	0.028	3
Geo_poor2	0.012	0.025	0.056	0.113	0.098	0.113	0.233	0.052	0.088	5
Geo_poor3	0.004	0.006	0.027	0.028	0.072	0.044	0.139	0.053	0.047	4
SimSLE	0.001	0.001	0.001	0.002	0.006	0.002	0.001	0.001	0.002	1

Fig. B2. L_1 (upper, in cm) and L_2 (lower, in cm^2) norms of model calibrations of 8 pumping tests for Case 1.

L1	Port 8	Port 11	Port 13	Port 15	Port 16	Port 18	Port 20	Port 23	Port 32	Port 35	Port 37	Port 38	Port 39	Port 40	Port 41	Port 42	Average	Rank
Geo_good	0.077	0.067	0.088	0.107	0.203	0.142	0.120	0.087	0.096	0.109	0.119	0.149	0.310	0.202	0.140	0.164	0.136	2
Geo_poor1	0.074	0.112	0.195	0.376	0.128	0.140	0.129	0.144	0.185	0.127	0.202	0.208	0.272	0.177	0.140	0.171	0.174	3
Geo_poor2	0.109	0.205	0.149	0.215	0.233	0.205	0.394	0.254	0.264	0.198	0.166	0.245	0.348	0.277	0.192	0.178	0.227	5
Geo_poor3	0.206	0.136	0.103	0.252	0.176	0.108	0.313	0.126	0.175	0.147	0.152	0.213	0.372	0.269	0.189	0.123	0.191	4
SimSLE	0.058	0.060	0.189	0.120	0.154	0.072	0.082	0.065	0.082	0.040	0.154	0.096	0.270	0.174	0.089	0.141	0.115	1

L2	Port 8	Port 11	Port 13	Port 15	Port 16	Port 18	Port 20	Port 23	Port 32	Port 35	Port 37	Port 38	Port 39	Port 40	Port 41	Port 42	Average	Rank
Geo_good	0.012	0.008	0.036	0.036	0.073	0.034	0.034	0.015	0.025	0.025	0.060	0.042	0.225	0.095	0.039	0.070	0.052	2
Geo_poor1	0.015	0.024	0.095	0.217	0.069	0.046	0.032	0.056	0.081	0.040	0.137	0.079	0.205	0.066	0.061	0.142	0.085	4
Geo_poor2	0.029	0.058	0.067	0.117	0.107	0.073	0.199	0.137	0.162	0.069	0.065	0.103	0.264	0.160	0.073	0.087	0.111	5
Geo_poor3	0.062	0.024	0.024	0.119	0.074	0.072	0.146	0.038	0.070	0.034	0.041	0.100	0.302	0.141	0.071	0.031	0.084	3
SimSLE	0.007	0.005	0.105	0.030	0.036	0.012	0.014	0.010	0.017	0.002	0.060	0.022	0.139	0.057	0.022	0.035	0.036	1

Fig. B3. L_1 (upper, in cm) and L_2 (lower, in cm^2) norms of model validations of 16 pumping tests for Case 1.

L1	Port 2	Port 5	Port 14	Port 17	Port 26	Port 29	Port 44	Port 47	Average	Rank
Geo-good					0.046	0.055	0.037	0.043	0.045	2
Geo-poor1					0.068	0.099	0.068	0.077	0.078	3
Geo-poor2					0.260	0.215	0.351	0.284	0.277	5
Geo-poor3					0.171	0.131	0.164	0.169	0.159	4
SimSLE					0.035	0.026	0.028	0.025	0.029	1

L2	Port 2	Port 5	Port 14	Port 17	Port 26	Port 29	Port 44	Port 47	Average	Rank
Geo-good					0.003	0.004	0.002	0.003	0.003	2
Geo-poor1					0.007	0.022	0.006	0.010	0.011	3
Geo-poor2					0.098	0.088	0.202	0.129	0.129	5
Geo-poor3					0.048	0.040	0.032	0.037	0.039	4
SimSLE					0.002	0.001	0.001	0.001	0.001	1

Fig. B4. L_1 (upper, in cm) and L_2 (lower, in cm^2) norms of model calibrations of 4 pumping tests for Case 2.

L1	Port 8	Port 11	Port 13	Port 15	Port 16	Port 18	Port 20	Port 23	Port 32	Port 35	Port 37	Port 38	Port 39	Port 40	Port 41	Port 42	Average	Rank
Geo-good	0.092	0.062	0.072	0.136	0.242	0.141	0.120	0.087	0.101	0.103	0.074	0.128	0.358	0.256	0.172	0.264	0.151	1
Geo-poor1	0.074	0.137	0.236	0.379	0.190	0.219	0.150	0.203	0.174	0.193	0.189	0.203	0.260	0.200	0.187	0.163	0.197	3
Geo-poor2	0.169	0.211	0.224	0.226	0.235	0.228	0.532	0.301	0.297	0.195	0.178	0.260	0.336	0.269	0.197	0.174	0.252	5
Geo-poor3	0.131	0.253	0.130	0.404	0.288	0.257	0.276	0.227	0.181	0.182	0.227	0.246	0.357	0.329	0.244	0.161	0.243	4
SimSLE	0.083	0.070	0.128	0.362	0.326	0.149	0.129	0.111	0.134	0.128	0.126	0.171	0.221	0.162	0.128	0.105	0.158	2

L2	Port 8	Port 11	Port 13	Port 15	Port 16	Port 18	Port 20	Port 23	Port 32	Port 35	Port 37	Port 38	Port 39	Port 40	Port 41	Port 42	Average	Rank
Geo-good	0.020	0.009	0.028	0.045	0.095	0.032	0.056	0.016	0.021	0.022	0.013	0.032	0.273	0.126	0.063	0.149	0.063	2
Geo-poor1	0.020	0.037	0.148	0.199	0.098	0.096	0.045	0.091	0.065	0.088	0.065	0.072	0.147	0.076	0.082	0.123	0.091	3
Geo-poor2	0.081	0.065	0.121	0.155	0.083	0.077	0.383	0.123	0.170	0.056	0.072	0.120	0.215	0.146	0.061	0.093	0.126	4
Geo-poor3	0.045	0.120	0.074	0.342	0.239	0.183	0.145	0.171	0.053	0.050	0.103	0.123	0.215	0.192	0.106	0.041	0.138	5
SimSLE	0.014	0.010	0.030	0.246	0.161	0.034	0.043	0.027	0.039	0.036	0.046	0.076	0.100	0.053	0.039	0.030	0.061	1

Fig. B5. L_1 (upper, in cm) and L_2 (lower, in cm^2) norms of model validations of 16 pumping tests for Case 2.

L1	Port 2	Port 5	Port 14	Port 17	Port 26	Port 29	Port 44	Port 47	Average	Rank
SimSLE*					0.035	0.026	0.028	0.025	0.029	2
SimSLE-Geo-good-Per					0.033	0.024	0.029	0.026	0.028	1
SimSLE-Geo-poor1-Per					0.034	0.025	0.029	0.028	0.029	3
SimSLE-Geo-poor2-Per					0.042	0.028	0.029	0.029	0.032	5
SimSLE-Geo-poor3-Per					0.035	0.028	0.030	0.025	0.029	4

L2	Port 2	Port 5	Port 14	Port 17	Port 26	Port 29	Port 44	Port 47	Average	Rank
SimSLE*					0.002	0.001	0.001	0.001	0.001	3
SimSLE-Geo-good-Per					0.002	0.001	0.001	0.001	0.001	1
SimSLE-Geo-poor1-Per					0.002	0.001	0.001	0.001	0.001	2
SimSLE-Geo-poor2-Per					0.002	0.001	0.001	0.001	0.002	5
SimSLE-Geo-poor3-Per					0.002	0.001	0.001	0.001	0.001	4

Fig. B6. L_1 (upper, in cm) and L_2 (lower, in cm^2) norms of model calibrations of 4 pumping tests for Case 3 (* denotes consistent results are used from Case 2 for comparison).

L1	Port 8	Port 11	Port 13	Port 15	Port 16	Port 18	Port 20	Port 23	Port 32	Port 35	Port 37	Port 38	Port 39	Port 40	Port 41	Port 42	Average	Rank
SimSLE*	0.083	0.070	0.128	0.362	0.327	0.149	0.129	0.111	0.134	0.128	0.126	0.171	0.221	0.163	0.128	0.105	0.158	5
SimSLE-Geo-good-Per	0.083	0.060	0.122	0.318	0.307	0.110	0.117	0.078	0.116	0.106	0.100	0.160	0.213	0.145	0.114	0.133	0.143	1
SimSLE-Geo-poor1-Per	0.085	0.065	0.169	0.339	0.228	0.116	0.131	0.093	0.114	0.094	0.122	0.174	0.196	0.138	0.110	0.145	0.145	2
SimSLE-Geo-poor2-Per	0.083	0.062	0.114	0.329	0.283	0.141	0.128	0.096	0.116	0.116	0.123	0.169	0.183	0.159	0.127	0.133	0.147	4
SimSLE-Geo-poor3-Per	0.079	0.065	0.100	0.367	0.315	0.079	0.118	0.094	0.121	0.105	0.106	0.158	0.221	0.157	0.125	0.126	0.146	3

L2	Port 8	Port 11	Port 13	Port 15	Port 16	Port 18	Port 20	Port 23	Port 32	Port 35	Port 37	Port 38	Port 39	Port 40	Port 41	Port 42	Average	Rank
SimSLE*	0.014	0.010	0.030	0.246	0.161	0.034	0.043	0.027	0.039	0.036	0.046	0.076	0.100	0.053	0.039	0.030	0.061	5
SimSLE-Geo-good-Per	0.014	0.007	0.065	0.177	0.148	0.022	0.040	0.014	0.034	0.034	0.036	0.059	0.087	0.049	0.035	0.054	0.055	3
SimSLE-Geo-poor1-Per	0.015	0.008	0.070	0.175	0.088	0.034	0.040	0.019	0.032	0.021	0.050	0.072	0.070	0.043	0.029	0.051	0.051	1
SimSLE-Geo-poor2-Per	0.014	0.007	0.029	0.189	0.124	0.035	0.040	0.021	0.032	0.034	0.044	0.073	0.069	0.049	0.038	0.054	0.053	2
SimSLE-Geo-poor3-Per	0.014	0.009	0.022	0.246	0.149	0.027	0.039	0.021	0.036	0.031	0.032	0.067	0.099	0.055	0.038	0.040	0.058	4

Fig. B7. L_1 (upper, in cm) and L_2 (lower, in cm^2) norms of model validations of 16 pumping tests for Case 3 (* denotes consistent results are used from Case 2 for comparison).

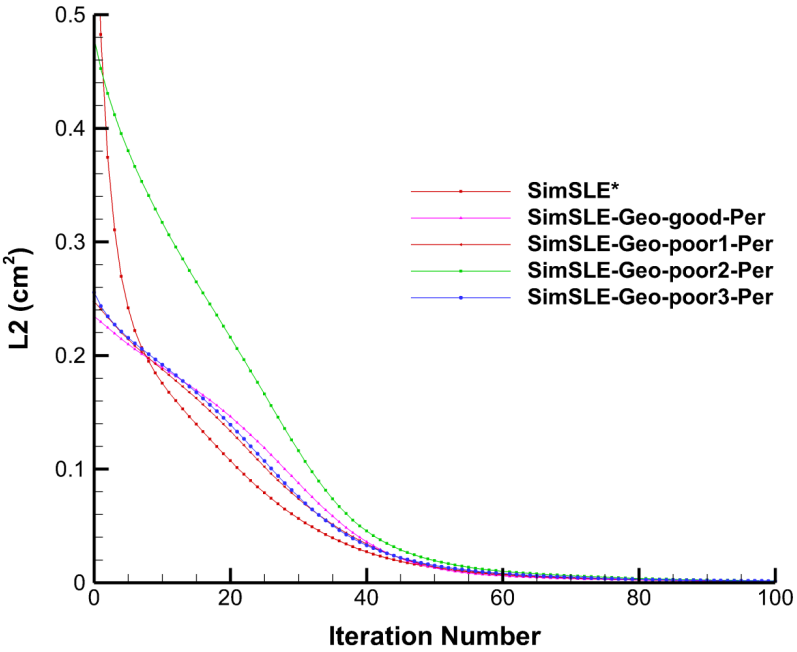


Fig. B8. L_2 (in cm^2) norms changes as a function of iteration number for calibration Case 3; (* denotes consistent results are used from Case 2 for comparison).

Supplementary Information C

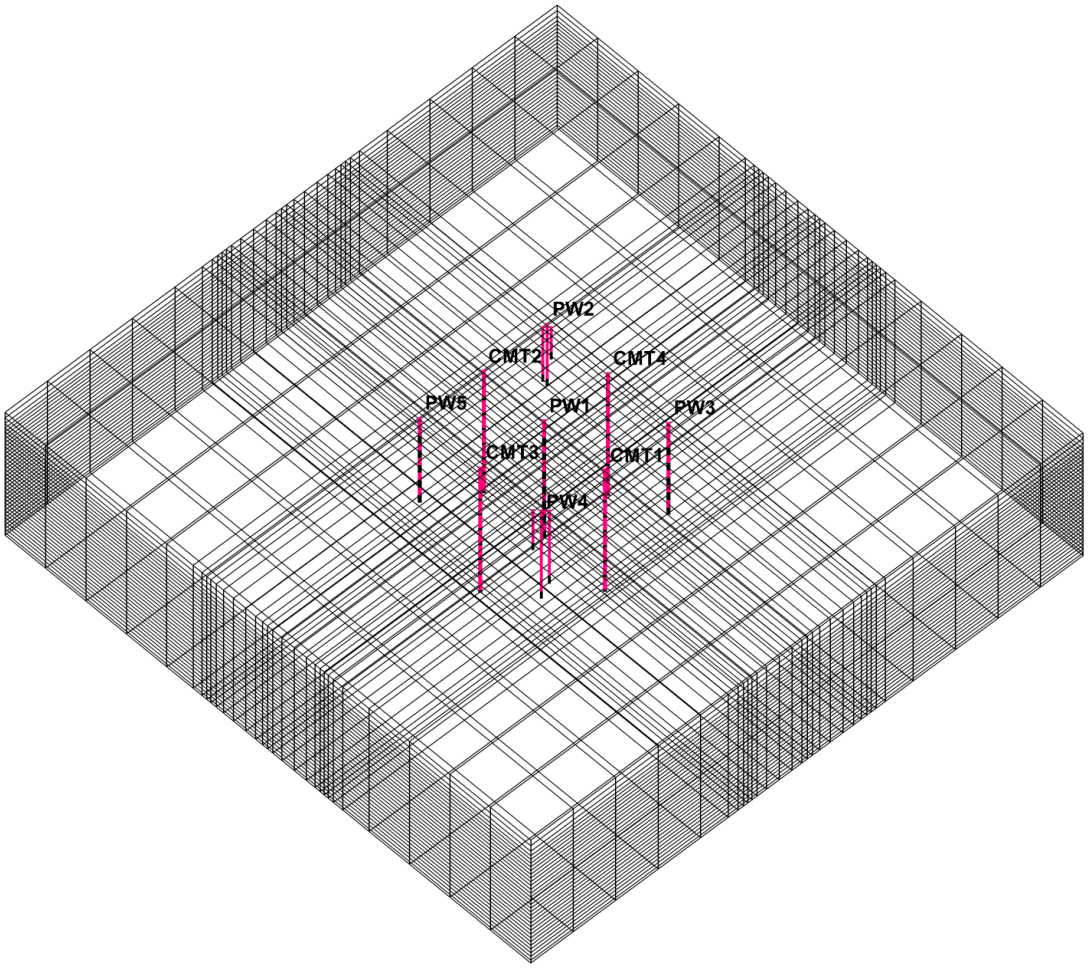


Figure C1: Computational grid for NCRS.

L1	Pw11	Pw14	Pw16	Pw23	Pw33	Pw53	Pw43	Average	Rank
Case 1a: Effective Value Model (isotropic)	0.17	0.17	0.10	0.07	0.14	0.17	0.15	0.14	**8**
Case 1b: Effective Value Model (anisotropic)	0.15	0.17	0.09	0.07	0.14	0.15	0.18	0.13	**7**
Case 2a: Geo-5	0.14	0.15	0.06	0.07	0.11	0.14	0.07	0.11	**6**
Case 2b: Geo-19	0.06	0.11	0.06	0.04	0.08	0.10	0.13	0.08	**5**
Case 3a: SimSLE_uniform	0.04	0.05	0.09	0.03	0.05	0.05	0.04	0.05	**4**
Case 3b: SimSLE_Geo-5	0.04	0.04	0.03	0.03	0.05	0.05	0.05	0.04	**3**
Case 3c: SimSLE_Geo-19	0.03	0.04	0.03	0.03	0.04	0.04	0.05	0.04	**1**
Case 3d: SimSLE_Permeater K	0.03	0.05	0.03	0.02	0.04	0.04	0.05	0.04	**2**

L2	Pw11	Pw14	Pw16	Pw23	Pw33	Pw53	Pw43	Average	Rank
Case 1a: Effective Value Model (isotropic)	0.05	0.10	0.03	0.01	0.04	0.07	0.04	0.05	**8**
Case 1b: Effective Value Model (anisotropic)	0.05	0.08	0.03	0.01	0.03	0.06	0.05	0.04	**7**
Case 2a: Geo-5	0.03	0.07	0.01	0.01	0.03	0.04	0.01	0.03	**6**
Case 2b: Geo-19	0.01	0.05	0.01	0.00	0.01	0.02	0.02	0.02	**5**
Case 3a: SimSLE_uniform	0.00	0.00	0.03	0.00	0.00	0.01	0.00	0.01	**4**
Case 3b: SimSLE_Geo-5	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	**3**
Case 3c: SimSLE_Geo-19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1**
Case 3d: SimSLE_Permeater K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**2**

Figure C2: L_1 (upper) and L_2 (lower) norms of model calibration of seven pumping tests.

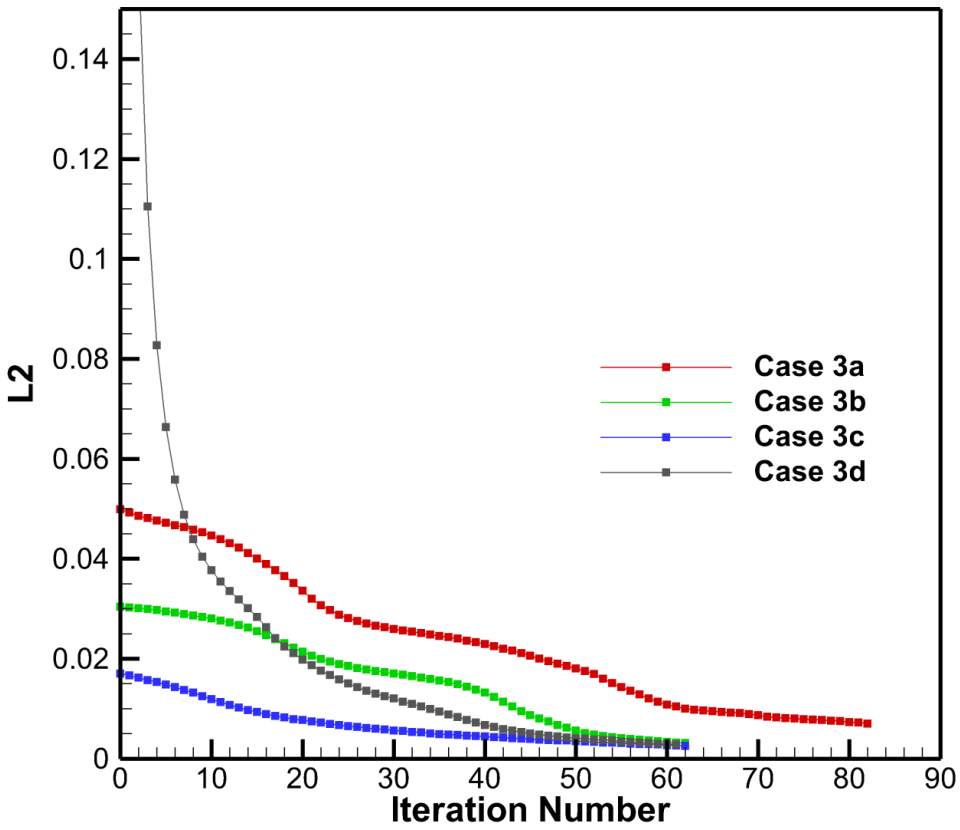


Figure C3: Mean square error (L_2 norm) as a function of iteration number.

L1	PW13	PW15	PW17	PW31	PW34	PW55	Average	Rank
Case 1a: Effective Value Model (isotropic)	0.22	0.05	0.04	0.07	0.03	0.09	0.08	**8**
Case 1b: Effective Value Model (anisotropic)	0.21	0.03	0.06	0.06	0.02	0.09	0.07	**6**
Case 2a: Geo-5	0.22	0.05	0.04	0.06	0.02	0.10	0.08	**7**
Case 2b: Geo-19	0.15	0.06	0.04	0.04	0.02	0.12	0.07	**5**
Case 3a: SimSLE_uniform	0.10	0.04	0.04	0.03	0.03	0.11	0.06	**4**
Case 3b: SimSLE_Geo-5	0.08	0.03	0.04	0.03	0.02	0.10	0.05	**2**
Case 3c: SimSLE_Geo-19	0.09	0.03	0.05	0.03	0.02	0.08	0.05	**3**
Case 3d: SimSLE_Permeater K	0.07	0.03	0.03	0.03	0.02	0.09	0.05	**1**

L2	PW13	PW15	PW17	PW31	PW34	PW55	Average	Rank
Case 1a: Effective Value Model (isotropic)	0.12	0.00	0.00	0.01	0.00	0.01	0.02	**8**
Case 1b: Effective Value Model (anisotropic)	0.09	0.00	0.00	0.01	0.00	0.01	0.02	**7**
Case 2a: Geo-5	0.08	0.00	0.00	0.01	0.00	0.02	0.02	**6**
Case 2b: Geo-19	0.04	0.01	0.00	0.00	0.00	0.02	0.01	**5**
Case 3a: SimSLE_uniform	0.02	0.00	0.00	0.00	0.00	0.02	0.01	**4**
Case 3b: SimSLE_Geo-5	0.01	0.00	0.00	0.00	0.00	0.01	0.01	**2**
Case 3c: SimSLE_Geo-19	0.02	0.00	0.00	0.00	0.00	0.01	0.01	**3**
Case 3d: SimSLE_Permeater K	0.01	0.00	0.00	0.00	0.00	0.01	0.01	**1**

Figure C4: L_1 (upper) and L_2 (lower) norms of model validations of seven pumping tests.