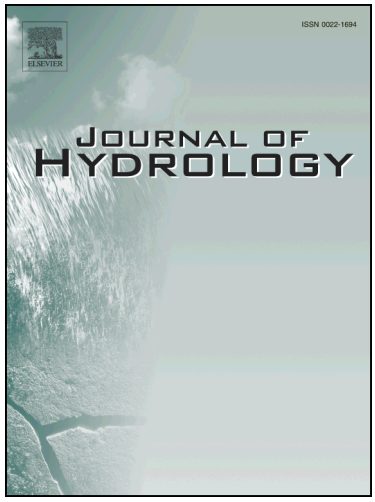
Accepted Manuscript



Research papers

Three-Dimensional Imaging of Aquifer and Aquitard Heterogeneity via Transient Hydraulic Tomography at a Highly Heterogeneous Field Site

Zhanfeng Zhao, Walter A. Illman

PII: S0022-1694(18)30100-8
DOI: https://doi.org/10.1016/j.jhydrol.2018.02.024
Reference: HYDROL 22573

To appear in: *Journal of Hydrology*

Received Date: 2 August 2017
Revised Date: 24 January 2018
Accepted Date: 12 February 2018

Please cite this article as: Zhao, Z., Illman, W.A., Three-Dimensional Imaging of Aquifer and Aquitard Heterogeneity via Transient Hydraulic Tomography at a Highly Heterogeneous Field Site, *Journal of Hydrology* (2018), doi: https://doi.org/10.1016/j.jhydrol.2018.02.024

Three-Dimensional Imaging of Aquifer and Aquitard Heterogeneity via Transient Hydraulic Tomography at a Highly Heterogeneous Field Site

Zhanfeng Zhao[1,2*], and Walter A. Illman[2]

For submission to *Journal of Hydrology*

January 24th, 2018

[1] Now at the Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

[2] Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, Ontario, Canada

[*] Corresponding author: Z. Zhao, zhaozhanfeng@igsnrr.ac.cn

Keywords: hydraulic tomography; subsurface heterogeneity; aquitard characterization; model calibration and validation; inverse modeling; model comparison

Abstract

Previous studies have shown that geostatistics-based transient hydraulic tomography (THT) is robust for subsurface heterogeneity characterization through the joint inverse modeling of multiple pumping tests. However, the hydraulic conductivity (K) and specific storage (S_s) estimates can be smooth or even erroneous for areas where pumping/observation densities are not high. This renders the imaging of interlayer and intralayer heterogeneity of highly contrasting materials including their unit boundaries difficult. In this study, we further test the performance of THT by utilizing existing and newly collected pumping test data of longer durations that showed drawdown responses in both aquifer and aquitard units at a field site underlain by a highly heterogeneous glaciofluvial deposit. The robust performance of the THT is highlighted through the comparison of different degrees of model parameterization including: (1) the effective parameter approach; (2) a geological zonation approach relying on borehole logs; and (3) a geostatistical inversion approach considering different prior information (with/without geological data). Results reveal that the simultaneous analysis of eight pumping tests with the geostatistical inverse model yields the best results in terms of model calibration and validation. We also find that the joint interpretation of long-term drawdown data from aquifer and aquitard units is necessary in mapping their full heterogeneous patterns including intralayer variabilities. Moreover, as geological data are included as prior information in the geostatistics-based THT analysis, the estimated K values increasingly reflect the vertical distribution patterns of permeameter-estimated K in both aquifer and aquitard

units. Finally, the comparison of various THT approaches reveals that differences in the estimated K and S_s tomograms results in significantly different transient drawdown predictions at observation ports.

1. Introduction

Investigations of groundwater problems have relied heavily on the accurate knowledge of the subsurface parameters such as hydraulic conductivity (K) and specific storage (S_s). Usually, these hydraulic properties are determined by fitting pumping test data from individual tests to analytical solutions with the assumption that the subsurface is homogeneous. Estimates of such traditional analyses can yield biased parameter estimates (e.g., *Wu et al.*, 2005; *Wen et al.*, 2010; *Illman et al.*, 2010; *Alexander et al.*, 2011; *Huang et al.*, 2011; *Berg and Illman*, 2011a,b; 2013, 2015) due to the fact that the subsurface is heterogeneous at multiple scales. Alternatively, small-scale measurements of cores, slug, single-hole, and flowmeter tests require a large number of measurements in order to adequately characterize subsurface heterogeneity. Thus, *Carrera et al.* (2005) advocated the more common use of inverse models to estimate hydraulic parameters based on observations of ambient hydraulic heads, artificial changes in heads, as well as other data such as solute concentration through either stochastic or deterministic groundwater flow and transport models (*Pool et al.*, 2015).

Among the various techniques to characterize the subsurface, pumping tests are considered to yield large-scale estimates of hydraulic parameters through the interpretation of drawdowns in both aquifer and aquitard layers. However, drawdown responses to such tests are only typically monitored in aquifers. In addition, as the tests are usually not run long enough, drawdown signals are detected only in high K zones connected to the pumped interval, leading to parameter estimates only for

aquifers (*Fogg and Zhang*, 2016; *Neuzil*, 1994, 1986). While the characterization of aquifers is important, the same applies to low K zones and aquitards (and their connectivity) for groundwater resource evaluation (e.g., *Scanlon et al.*, 2003; *Konikow and Neuzil*, 2007), solute transport (e.g., *Johnson et al.*, 1989; *Hendry and Wassenaar*, 2000), land subsidence (*Galloway and Burbey*, 2011; *Zhuang et al.*, 2017a, b) and also waste storage (*Montazer and Wilson*, 1984; *Neuzil*, 1994).

Aquitard hydraulic parameters are typically obtained by performing laboratory permeameter tests together with consolidation tests on small-scale samples (e.g., *Keller et al.*, 1989; *Alexander et al.*, 2011). However, laboratory tests can yield hydraulic parameter estimates that may not be representative of field conditions. As reviewed by *van der Kamp* (2001), field testing methods for aquitard characterization include the analysis of slug and pumping tests (e.g., *Neuman and Witherspoon*, 1972; *Keller et al.*, 1989), monitoring of pore-pressure changes and settlement due to surface loading (*van der Kamp and Maathuis*, 1985), as well as monitoring of seasonal head fluctuations (*Davis*, 1972; *Keller et al.*, 1989).

van der Kamp (2001) noted that the most common methods provide only a one-dimensional value of hydraulic diffusivity ($\alpha = K_v/S_s$) of the aquitard, where K_v = vertical hydraulic conductivity and a value of K_v is inferred by relying on a laboratory estimate of S_s. He also noted that laboratory estimates of S_s are typically several orders of magnitude larger than those estimated in the field. The commingled use of field and laboratory values can be problematic and can result in inaccurate estimates of K_v.

Other information, such as land subsidence data recorded by multiple vertical extensometers (e.g., *Cleveland et al*., 1992; *Zhuang et al*., 2015, 2017b), could also be analyzed to determine aquitard hydraulic parameters. However, despite its importance, very little work has been done to map the variability of hydraulic parameters in aquitards. This may partly be a result of significant difficulties in obtaining even a single reliable hydraulic parameter estimate of aquitard units, notwithstanding their heterogeneities. Clearly, new methods are necessary for obtaining more reliable hydraulic parameter estimates within aquitards and mapping heterogeneities within them.

Hydraulic tomography (HT) has become a mature inverse modelling technique to map both high and low *K* features of the subsurface. It relies on the joint inverse modeling of multiple sets of pressure heads that are obtained from different observation intervals, while pumping and/or injecting water at different locations of target aquifers. Thus far, HT has been investigated by several research groups through synthetic studies (e.g., *Yeh and Liu*, 2000; *Bohling et al.*, 2002; *Zhu and Yeh*, 2005, 2006; *Cardiff et al.*, 2009, 2013; *Liu and Kitanidis*, 2011; *Mao et al.*, 2013), laboratory experiments (e.g., *Liu et al.*, 2002, 2007; *Brauchler et al.*, 2003; *Illman et al.*, 2007, 2008, 2010; *Berg and Illman*, 2011a; *Liu and Kitanidis*, 2011; *Zhao et al.*, 2015; *Zhou et al.*, 2016) and testing at the field-scale (*Bohling et al.*, 2007; *Cardiff et al.*, 2009, 2012; *Illman et al.*, 2009; *Berg and Illman*, 2011b, 2013, 2015; *Brauchler et al.*, 2011; *Hochstetler et al.*, 2016; *Paradis et al*., 2016; *Zhao and Illman*, 2017). Compared to traditional aquifer characterization approaches, HT has shown to be

superior in recovering the heterogeneity as well as predicting independent pumping tests not used for model calibration (*Illman et al.*, 2010, 2015; *Berg and Illman*, 2011a,b, 2015).

In order to interpret multiple pumping and observation data from HT tests, different inverse modeling methods have been developed. Two most frequently used inverse modeling algorithms are the quasi-linear geostatistical approach (QL) (*Kitanidis*, 1995) and the successive linear estimator (SLE) (*Yeh et al*., 1996), both of which rely on the cross-covariance between hydraulic head and aquifer parameters (e.g., *Yeh and Liu*, 2000; *Zhu and Yeh*, 2005; *Illman et al*., 2009; *Berg and Illman*, 2011a,b; *Cardiff and Barrash*, 2011; *Mao et al*., 2013). Alternatively, the Kalman filter (KF) as well as its low rank versions, such as the ensemble KF (e.g., *Nowak*, 2009; *Li et al.*, 2012; *Schöniger et al.*, 2012), generalized compressed state KF (e.g., *Kitanidis*, 2015; *Li et al*., 2015), and the spectral KF (e.g., *Ghorbanidehno et al.*, 2015, 2017), have also been increasingly used for hydraulic parameter estimation, through a Bayesian framework to continuously assimilate hydraulic head and/or concentration data.

For the majority of the above HT studies, geostatistics forms the backbone of various inverse methods. While the geostatistics-based HT offers many advantages (*Yeh and Šimůnek*, 2002), it could produce overly smooth distributions of subsurface heterogeneity, when only few pumping tests and monitoring data are available (*Yeh and Liu*, 2000; *Cardiff et al*., 2013; *Illman et al*., 2015; *Zhao and Illman*, 2017). For example, in the field studies at the North Campus Research Site (NCRS) located on

the University of Waterloo campus in Waterloo, Ontario, Canada, both the THT (*Berg and Illman*, 2011b) and steady state hydraulic tomography (SSHT) (*Berg and Illman*, 2013) analyses of four pumping tests captured the most salient heterogeneity patterns of the highly heterogeneous glaciofluvial deposits. However, despite the relatively large number of monitoring intervals installed within the well field, stratigraphic boundaries between aquifer and aquitard units were ambiguous. More importantly, high *K* values were estimated for the lower portion of the model domain, where the known geology indicates the presence of an aquitard consisting of clay. *Berg and Illman* (2011b, 2013) pointed out that little to no drawdown responses were observed in monitoring wells located in the low *K* material during the relatively short-duration pumping tests, thus leading to unsatisfactory results. Later, *Berg and Illman* (2015) were able to map those low *K* zones, but only after using permeameter *K* values for conditioning. Therefore, one could question whether the geostatistics-based inversion approach could be reliably used to map unit boundaries and more importantly, aquitard units consisting of low *K* materials with pumping test data alone or not.

Thus far, HT studies have not been previously conducted to map both aquifer and aquitard units and their connectivity. Because geostatistics-based HT yields smoothed hydraulic parameter distributions when data are sparse, the inclusion of other types of information becomes necessary to map the various units (i.e., interlayer heterogeneity) and the heterogeneity within them (i.e., intralayer heterogeneity). Lessons from our previous studies (*Berg and Illman*, 2011, 2013) also point to the

need for longer pumping tests to stress the aquitard units to characterize them with pumping tests alone.

In order to overcome the issue of smooth hydraulic parameter estimates of the geostatistics-based inversion approaches, the importance of geological data was investigated for HT analyses. In particular, *Zhou et al*. (2014) proposed an image-guided method which extracts the structure information from the cross sections of geology and incorporating it in the inversion process. Later, *Zhou et al*. (2016) extended this method by including the Markov-chain Monte Carlo sampler to update and select the most plausible geological models for the inverse problem. This image-guided inversion approach was then tested in synthetic HT studies by Soueid Ahmed et al. (2015), but not through laboratory sandbox or field experiments.

On the other hand, *Illman et al*. (2015) compared the SSHT analyses based on geostatistical inversion to those based on effective parameter and geological models, to test whether subsurface conceptualizations of the *K* structure at lower resolutions for a sandbox aquifer could yield similar results to the geostatistical inverse modeling approach or not. They found that the geostatistical inversion approach performed best in terms of model calibration and validation, but the geological model with perfectly known stratigraphy came a close second. Then, *Zhao et al*. (2016) evaluated the performances of geological zonation models of varying accuracy for SSHT through the same laboratory sandbox, and *Luo et al*. (2017) extended the work of *Zhao et al*. (2016) to the THT case. Results from the sandbox studies revealed that, both accurate and inaccurate geological models could be well calibrated, despite the estimated *K*

values for the poor geological models being quite different from the actual values. These studies also concluded that: (1) using a geological model as prior mean K distributions in geostatistical inverse models resulted in the preservation of geological features, especially in areas where drawdown data were not available; and (2) transient inversions are necessary by treating both K and S_s to be heterogeneous to jointly obtain reliable K and S_s estimates for making accurate predictions of transient drawdown events. The findings by *Zhao et al.* (2016) and *Luo et al*. (2017) were based on experiments conducted in a laboratory synthetic aquifer constructed with various types of sands, which did not contain low K materials (e.g., clay, silty clay) as encountered at the NCRS *(Alexander et al*., 2011; *Berg and Illman*, 2011b). Therefore, efforts other than *Berg and Illman* (2011b, 2013, 2015) are needed to examine whether the HT approach can map aquitard units consisting of low K materials through pumping tests alone.

Most recently, *Zhao and Illman* (2017) performed a new SSHT study at the NCRS. The unique contribution of *Zhao and Illman* (2017) was that they tried to stress the low K layers and included both steady and quasi-steady state drawdown data from low K zones into the SSHT analysis. Compared to the previous HT analyses of *Berg and Illman*. (2011b, 2013) in which the bottom aquitard layer was incorrectly mapped as a high K zone, slight improvements in the characterization of the lower aquitard were obtained by *Zhao and Illman* (2017). Yet, the consistency between the estimated and permeameter test K values was still poor in silt and clay layers, due to the fact that only late time pressure heads indicating steady or quasi-steady state were

selected for model calibration. However, when transient data are available from aquitard units, more complete information could then be utilized for THT analysis.

Following this line of thought, the first purpose of this study is to perform THT analysis using long-term transient drawdowns obtained from both aquifer and aquitard units to investigate whether more accurate maps of both units as well as their intralayer heterogeneity could be obtained through the inversion of pumping test data alone.

Furthermore, the importance of geological data for THT has not been investigated rigorously in the field. In order to examine these issues, the second purpose of the study is to extend the work of *Zhao and Illman* (2017) to the transient case and compare the model calibration and validation performances among several approaches: (1) the effective parameter approach by treating the site to be homogeneous, both isotropic and anisotropic; (2) the geological zonation approach treating each layer to be homogeneous; and (3) the highly parameterized geostatistical inversion approach.

2. Data Used for Analysis

2.1 Site Description and Geology

Data for this THT study has been collected at the North Campus Research Site (NCRS) located on the University of Waterloo (UW) campus in Waterloo, Ontario, Canada. Previous investigations revealed that the near surface is highly heterogeneous and composed of multiple layers of glacial tills (*Alexander et al.*, 2011; *Karrow*,

1979; *Sebol*, 2000). Based on the continuous core samples of a 30-m deep borehole, the geology beneath the NCRS consists of, from younger to older age, the Tavistock Till, Maryhill Till and Catfish Creek Till (*Karrow*, 1979). The surface layer in our study area is recognized as the Maryhill Till, which consists mainly of silty clay accompanied with few stones, while the overlying Tavistock Till only exists as erosional remnants at the site. The Catfish Creek Till is composed of stiff stony silt to sandy silt, which is hard and difficult to drill, thus is treated to be the base of our groundwater models (*Berg and Illman*, 2011b).

The main aquifer zone of the NCRS, composed of high *K* sand to sandy gravel, is located from 8 to 13 m below ground surface (mbgs). Detailed descriptions of core samples indicate that the aquifer zone consists of two high *K* units separated by a discontinuous low *K* layer. The thin aquitard layer separating the two aquifers is discontinuous and is known to contain stratigraphic windows allowing for hydraulic connection (*Alexander et al.*, 2011). Situated below and above the main aquifer zone are aquitard layers composed of low *K* silts and clays. The overlying aquitard layer also is known to contain stratigraphic windows at various locations (*Martin and Frind*, 1998). Previous pumping tests performed at the site (*Alexander et al.*, 2011) indicated that the aquifer at the NCRS behaves as a confined to semi-confined system.

2.2 Field Data for Building Groundwater Models

2.2.1 Collection of core samples and laboratory analyses

In previous studies by *Alexander et al*. (2011) and *Berg and Illman* (2011b), four continuous multichannel tubing wells (CMT1, CMT2, CMT3, CMT4), three multi-screened wells (PW1, PW3, PW5) and two well clusters (PW2, PW4) were installed in an area of 15 m by 15 m at the NCRS. Fig. 1a is a plan view showing well locations, while Fig. 1b provides a three-dimensional perspective view of wells, corresponding pumping and observation intervals, as well as intervals sealed with bentonite at the site.

Each CMT well has seven 10-cm long screens and the screens are spaced 2 m apart. The upper most screens are installed between 4.5 and 5.5 mbgs, and the deepest screens are placed from 16.5 to 17.5 mbgs. PW1 is completed to an approximate depth of 18 m and screens are placed at eight different elevations. PW3 and PW5 are multi-screened at five different elevations and extends approximately to 12 mbgs. PW2 and PW4 are well clusters that each consists of three separate wells and screened over a 1 m interval. Screen elevations for PW2 are 4, 7, and 8 mbgs, while screen elevations for PW4 are 5, 8.5, and 11.5 mbgs.

During the drilling and installation of all CMT and PW wells, continuous cores were collected to characterize the site geology. After splitting the core into half along its length, soil texture was classified based on the layering observed at the scale of the core. Then, samples were extracted at 10 or 50 cm intervals for laboratory falling head permeameter tests. Specifically, core samples from five wells (CMT1, CMT2, CMT3,

CMT4 and PW1) were initially tested by *Alexander et al.* (2011) and samples from the other four wells (PW2, PW3, PW4, PW5) were tested by *Zhao and Illman* (2017). These *K* values were used in later sections to quantitatively and qualitatively compare the *K* estimates from different approaches along each borehole.

The borehole logs of the above nine wells, together with additional nine wells summarized from previous work by *Sebol* (2000), were compiled to construct a geological model for the NCRS. Fig. 1a shows the distribution of wells from which geological information was obtained. Based on the soil types and corresponding depth information, 19 different layers representing seven different material types were defined along all boreholes.

2.2.2 Description of pumping tests

To date, a total of 15 pumping/injection tests have been conducted at the NCRS to stress the multiple aquifer-aquitard system in a tomographic fashion. We have summarized the details to the pumping/injection tests in Table S1. 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) for previous HT analyses (*Berg and Illman*, 2011, 2013, 2015). These pumping tests mainly stressed the aquifer layers of the NCRS, since groundwater can be readily pumped from these units. Although pressure transducers were installed in observation ports located in aquitard layers, no drawdown responses have been observed by *Berg and Illman* (2011b) from the bottom ports consisting of Catfish Creek Till. *Berg and Illman* (2013) suggested that

it may take several days to even weeks to induce observable drawdowns into the lower aquitard zones of the system, when pumping from the aquifer zones. Therefore, to obtain more complete drawdown information, six additional pumping and injection tests were conducted by *Zhao and Illman* (2017) to directly stress the aquitard zones at PW1-1, PW1-6, PW1-7, PW2-3, PW3-1, and PW5-1. Due to the low permeable nature of surrounding deposits, pumping and injection tests at these six well locations were conducted at flow rates that were generally no more than 2.0 L/min, as shown in Table S1. Noticeable drawdowns (> 0.1m) were generally observed within 6.5 hours from the ports whose elevations were similar to the pumping port, except for the test at PW1-7 which lasted up to 26.5 hours. In particular, the PW1-6 and PW1-7 intervals could only be pumped at approximately 1.0 L/min. Such low flow rates have induced measureable drawdowns at the bottom intervals within the aquitard zone (i.e., CMT -6 and -7 ports), while no drawdown has been observed from upper intervals (i.e., CMT -1, -2, -3, -4 and -5 ports).

During each pumping/injection test, pressure transducers were placed at all available observation intervals. Specifically, all CMT wells from the top to bottom intervals (e.g., CMT1-1 to CMT1-7) were instrumented with 0 - 15 psig (model MP100: Micron Systems) pressure transducers, while the bottom intervals (e.g., CMT1-7, CMT2-7, CMT3-7, and CMT4-7) were monitored with an electronic water level tape by *Berg and Illman* (2011b). PW2 and PW4 wells were monitored with 0 - 5 or 0 - 10 psig pressure transducers (model 3001 LT Leveloggers Junior: Solinst). FLUTe water systems (FLUTe Ltd.) were installed in the multi-screened pumping

wells (e.g., PW1, PW3, and PW5) during pumping/injection tests. A blank FLUTe liner was installed, when the wells were not being pumped. Each FLUTe water system contained five vented pressure transducers (Level Troll: In Situ) that were designed to fit the screened intervals of PW1, PW3, and PW5. The FLUTe systems seal off the entire well to prevent short circuiting of pressure across the multiple open screened intervals.

We utilized data from 12 pumping/injection tests to perform the THT analysis at the NCRS. Data from eight tests (PW1-1, PW1-4, PW1-6, PW1-7, PW2-3, PW3-3, PW4-3, and PW5-3) were selected for model calibration, while the other four tests (PW1-3, PW1-5, PW5-4, and PW5-5) were reserved for model validation. Test data from PW3-1, PW3-4, and PW5-1 were not selected due to the fact that drawdowns were insignificant or only observed at very few ports. In addition, test data affected by the Noordbergum effect (*Verruijt*, 1969; *Berg et al.*, 2011, 2015) were not included in the analysis.

For the data recorded by the transducers in the CMT wells, a 10-point centrally weighted moving average was applied to remove the static sensor noise. Since the data were collected at a high frequency (4 Hz at early time and 1 Hz at late time), the application of this filter did not significantly impact the shape of the drawdown curve. Three to five points were selected manually from the early, intermediate, and late time of each transient drawdown curve. In total, we selected 522 pressure head data for model calibration and used 348 head data for model validation.

3. Description of Models Used for Hydraulic Tomography Analysis

We compared three approaches different in model conceptualizations and complexities for HT analysis namely, (1) the effective parameter approach, both isotropic and anisotropic; (2) the geological zonation approach; and (3) the highly parameterized geostatistical approach.

In order to simulate transient groundwater flow for all models, a three-dimensional model with dimensions of 70 m × 70 m × 17 m was constructed and discretized into 31,713 variably-sized rectangular finite elements. We note that the groundwater flow model did not consider poroelastic effects.

This model was larger than the 45 m × 45 m × 15 m model used by *Berg and Illman* (2011b, 2013, and 2015). Such a new domain, on one hand, was designed to incorporate additional borehole logs for constructing a site geological model. On the other hand, this larger model minimizes the impacts of boundary conditions, when including new pumping and injection tests of longer durations.

The elements were gradually refined from the model boundary to the central 15 m × 15 m × 17 m well clustering area, with the block size decreasing from 5 m × 5 m × 0.5 m to 0.5 m × 0.5 m × 0.5 m. The top and bottom faces were defined as no-flow boundaries due to the presence of low K units at the top and bottom areas of the modeling domain, while constant heads were assigned to the remaining boundaries.

3.1 Case 1: Effective Parameter Approach

We first estimated the effective K and S_s values for the multiple aquifer-aquitard system, by coupling the groundwater flow model HydroGeoSphere (HGS) *(Therrien et al*., 2005) with the parameter estimation code, PEST (*Doherty*, 2005). Two cases were considered in this approach. Case 1a treated the aquifer-aquitard system as homogeneous/isotropic, in which only the effective K and S_s values were estimated. Case 1b treated the system as being homogeneous/anisotropic and the effective K_x, K_y, K_z and S_s values were estimated. The initial values of K and $K_x/K_y/K_z$ were 8.0×10^{-6} m/s, with a minimum bound of 1.0×10^{-10} m/s and a maximum bound of 1.0×10^{-1} m/s. The initial value of S_s in both Case 1a and 1b was 1.0×10^{-4} /m with minimum and maximum bounds of 1.0×10^{-8} /m and 1.0×10^{-3} /m, respectively. These initial values were geometric means of individual K and S_s estimates obtained by *Berg and Illman* (2011b) through matching the transient drawdown curve at each observation port for a pumping test conducted at PW1-3.

3.2 Case 2: Geological Zonation Approach

As introduced previously, borehole logs of 18 wells completed to different depths were compiled based on determined soil type information. In total, 19 different layers representing seven different material types were defined along all boreholes. To investigate the value of geological data for THT data interpretation, we constructed a three-dimensional geological model with dimensions of 70 m × 70 m × 17 m using a commercial software Leapfrog Hydro (*ARANZ Geo. Limited*, 2015). Leapfrog Hydro

utilizes the Fast Radial Basis Function method, which is an effective way of implementing dual kriging to interpolate stratigraphy between boreholes based on the known geological layering information from available wells. Fig. 2 shows four cross-sections (A-A′, B-B′, C-C′, and D-D′, as indicated in Fig. 1a) extracted along different directions among the central nine wells to illustrate the interpolated geological layers. We also present the locations of wells and screens for cross-sections C-C′ and D-D′.

Based on the interpolated stratigraphy, two geological models were built: (1) a 5-layer geological model (Case 2a), constructed by merging some layers with low K material, specifically layers 1 through 10 as layer 1*, layers 12 through 14 as layer 12*, and layers 16 through 19 as layer 16*; and (2) a 19-layer geological model (Case 2b) to take full advantage of the available stratigraphy information. Both geological models were then utilized to create groundwater flow models using the same grid as described earlier.

Calibrations of the 5-layer (Case 2a) and the 19-layer (Case 2b) geological models were also performed by coupling HGS with PEST. For both model calibrations, the initial K value was set as 8.0×10^{-6} m/s for all layers, with a minimum bound of 1.0×10^{-10} m/s and a maximum bound of 1.0×10^{-1} m/s, while the initial S_s value was 1.0×10^{-4} /m with minimum and maximum bounds of 1.0×10^{-8} /m and 1.0×10^{-2} /m, respectively. In both Cases 2a and 2b, the estimated parameters were treated to be uniform and isotopic in each layer.

3.3 Case 3: Geostatistical Inversion Approach

We analyzed the eight pumping tests using the geostatistics-based Simultaneous Successive Linear Estimator (SimSLE) code developed by *Xiang et al.* (2009). SimSLE inverts all the data sets simultaneously, thus providing more constraints to the inverse problem (*Xiang et al.*, 2009) compared to the sequential inversion approach (*Yeh and Liu*, 2000). The model domain used for the SimSLE is identical to the one introduced for the other models and we assume that each element is isotropic during the estimation process.

In SimSLE, natural log values of hydraulic parameters (i.e., *ln K* and *ln* S_s) are treated as a stochastic process, and the corresponding unconditional means, spatial covariance functions and structure parameters (correlation scales λ_x, λ_y, λ_z and the variances, σ_{lnK}^2, σ_{lnSs}^2) of hydraulic parameters are assumed to be known *a priori*. The inversion process starts with cokriging using available measurements of hydraulic property and pressure heads to produce the conditional property field. The stochastic conditional means of these parameters are used for predictions of pressure heads at observation ports. The cokriged parameter field is then iteratively updated by SimSLE to minimize the differences between observed and simulated heads.

In this study, the exponential covariance model is adopted for the parameter fields. The initial correlation scales of the *K*- and S_s- fields are assumed as $\lambda_x = \lambda_y = 4$ m, $\lambda_z = 0.5$ m, and the variances are set to $\sigma_{lnK}^2 = \sigma_{lnSs}^2 = 5.0$, which are the values used in previous HT studies at the NCRS (*Berg and Illman*, 2011b, 2013; *Zhao and*

Illman, 2017). Other initial inputs to the inverse model include initial guesses for the K and S_s fields.

Zhao et al. (2016) and *Luo et al.* (2017), through laboratory sandbox experiments as well as *Zhao and Illman* (2017) through their field study found that the geostatistical inversion approach using a geological model as prior information preserved geological features where drawdown measurements were lacking compared to using a homogenous K as a prior. Therefore, we considered four scenarios (Cases 3a, 3b, 3c and 3d) different in the initial guesses of K for the geostatistical inversion approach to meet our study purposes. In Case 3a, the inversion starts with homogenous mean fields of $K = 8.0 \times 10^{-6}$ m/s and $S_s = 1.0 \times 10^{-4}$ m^{-1}, which are the same as the initial values used in the effective parameter and geological zonation approaches. For the other three cases (Cases 3b – 3d), heterogeneous mean K fields based on geological zonations are used as prior information. Specifically, inverse estimations of K and S_s distributions in Cases 3b and 3c start with the estimated K values from model calibrations of Cases 2a and 2b, respectively. In Case 3d, instead of starting from the calibrated K fields, we use the 19-layer geological model populated with permeameter tested K values as the prior mean K distribution. These permeameter test K values are geometric means calculated from laboratory measurements of soil samples located in the same layer of the geological model, shown in Table S2 (Supplementary Information section). For the layers that have no core sample data, measurements of similar soil material are assigned.

4. Results and Discussions

4.1 Model Calibration Results

THT analysis of eight tests for the effective parameter and geological zonation approaches were performed on a PC with a quad-core CPU and 24 GB of RAM. The computational time increased with model complexity. Specifically, Case 1a took less than 2.5 hours and Case 1b took approximately 4.5 hours. Calibration of the 5-layer geological model (Case 2a) took approximately 11.5 hours, while the calibration of the 19-layer geological model (Case 2b) was completed in seven days after 1,075 PEST model calls to estimate 38 unknowns. Note here that each ‘‘model call’’ consisted of transient forward simulation of eight tests and PEST optimization sequentially, which took about 9.5 minutes using a single CPU. This long computational time could have been reduced if a parallel computing environment was implemented. On the other hand, geostatistical inversions (Cases 3a – 3d) using SimSLE were performed on a PC-cluster using 16 processors with 192 GB of RAM and all cases of the geostatistical inversion approach converged within two days.

4.1.1 Case 1: Effective parameter approach

The estimated K and S_s values as well as their corresponding posterior 95% confidence intervals of the effective parameter models are summarized in Table 1. Examination of Table 1 reveals that, when treating the medium to be homogeneous/ isotropic, Case 1a yields a K value of 2.38×10^{-5} m/s and a S_s value of 9.34×10^{-6}/m. For the homogeneous/anisotropic case (Case 1b), K_x, K_y, and K_z are estimated as 1.85

$\times 10^{-5}$ m/s, 2.55×10^{-5} m/s, and 3.77×10^{-7} m/s, respectively, and S_s is estimated as 1.39×10^{-5} /m.

When the medium is treated to be homogeneous, the estimated parameters are found to vary with the observation and pumping locations (e.g., *Wen and Chen*, 2006; *Liu et al.*, 2007; *Huang et al.*, 2011). By individually calibrating one anisotropic effective parameter model to four pumping test, *Berg and Illman* (2015) estimated that K_x ranged between 4.0×10^{-6} and 2.9×10^{-5} m/s depending on the pumping location, K_y ranged between 4.8×10^{-6} and 7.3×10^{-5} m/s, K_z ranged between 3.0×10^{-8} and 1.0×10^{-6} m/s, and S_s ranged between 1.0×10^{-7} and 6.8×10^{-4} /m.

By performing grain size analysis of 270 core samples and permeameter test analyses for 471 core samples, *Alexander et al*. (2011) obtained K estimates ranging between 3.2×10^{-11} m/s to 2.5×10^{-3} m/s and from 5.8×10^{-10} m/s to 2.8×10^{-4} m/s, respectively. Our estimates of K and S_s fall within the range of these previous studies, suggesting the validity of the estimated values for Cases 1a and 1b.

In a recent SSHT study at the NCRS, *Zhao and Illman* (2017) simultaneously calibrated the same effective parameter models, but used steady state and quasi-steady state data from seven pumping tests. The estimated effective K value was 8.43×10^{-6} m/s for the isotropic case. For their anisotropic case, K_x, K_y, K_z were estimated as 1.04×10^{-5} m/s, 1.19×10^{-5} m/s and 6.37×10^{-7} m/s, respectively.

Through controlled sandbox studies, *Illman et al*. (2015) concluded that, when calibrated to a large number of observation data from multiple pumping tests instead of data from an individual pumping test, the effective groundwater model yields

improved drawdown predictions for pumping tests not used in the calibration effort. *Yeh et al*. (2015) also stated that, in order to predict the average spatial trend of observed heads in a heterogeneous aquifer, many head measurements distributed in the aquifer must be used so that the effective hydraulic properties of the equivalent homogeneous groundwater model could be determined. Compared to all previous HT studies at the NCRS, the THT analysis performed in this study included transient data from additional pumping tests performed in aquitard layers, instead of selecting only one data point from each curve for the steady state case (*Zhao and Illman*, 2017) or transient data only from individual pumping tests *(Berg and Illman*, 2015). Consequently, the differences between the effective values of the current study and those of previous studies could be attributed to the inclusion of more transient drawdown data. Meanwhile, we can expect that the estimated K and S_s values to be more representative of the heterogeneous properties of the multiple aquifer-aquitard system in an averaged sense than those obtained from previous HT studies.

4.1.2 Case 2: Geological modeling approach

Calibrations of the geological models are performed by treating each layer to be homogeneous and isotropic. The estimated K and S_s distributions are plotted in Figs. 3a and 4a for the 5-layer model, and in Figs. 3b and 4b for the 19-layer model, respectively. The estimated values and their 95% confidence intervals are summarized in Tables 2 and 3.

As previously noted, the main aquifer zone of the NCRS consists of two high K units separated by a thin discontinuous aquitard layer. Situated below and above the

aquifer zone are aquitard layers composed mainly of low *K* silts and clays. Calibration results of the 5-layer geological model (Case 2a; Fig. 3a and Table 2) reveal that a high *K* value is estimated for the sand and gravel layer 15 of the aquifer zone in the middle model domain, while relatively low *K* values are obtained for the merged aquitard layers 12* and 16*, which consist of low permeable silt and clay. For the merged layer 1* containing both low permeable silt and clay and high permeable sand layers, the *K* estimate is close to the initial value of 8.0×10^{-6} m/s, which could be a result of using only one geological layer to represent multiple soil types. However, sand layer 11, expected to have a high *K*, has the lowest *K* value among the five layers as shown in Fig. 3a, which is inconsistent with geological data. Through a controlled laboratory sandbox study, *Zhao et al*. (2016) showed that the estimated *K* values for some layers of a simplified geological model by merging layers of similar material can also be inconsistent with permeameter test values. Therefore, this kind of inconsistency could potentially result through the use of a less complex geological model for the NCRS aquifer and aquitard system which contains both interlayer and intralayer heterogeneity.

When the 19-layer geological model is used for model calibration, we see from Fig. 3b and Table 3 that, although the general distribution pattern of *K* is similar to Fig. 3a, there are more variations in *K* values compared to the 5-layer model. Specifically, in the upper part of model domain, low *K* layers are more clearly recovered in Fig. 3b.

The estimated S_s values are found to vary in the range of 2.29×10^{-7} /m and 9.28×10^{-6} /m for the 5-layer geological model, while for the 19-layer geological model, S_s varies between 1.02×10^{-6} /m and 5.88×10^{-3} /m. In Fig. 4a, S_s estimates obtained by calibrating the 5-layer geological model show a pattern of an obviously low value assigned to the top portion of the domain, which may have resulted by merging layers 1 through 10. In contrast, the S_s distribution estimated by calibrating the 19-layer geological model (Case 2b) shows a pattern with high S_s values at the top and bottom areas and relatively low values in the central domain. More specifically, the estimated S_s values for layers 17 and 18 are one order of magnitude higher than S_s estimates of the other layers. During the previous site characterization effort at the NCRS, *Alexander et al*. (*2*011) obtained a range of S_s estimates between 2.6×10^{-8} /m and 3.8×10^{-3} /m with a geometric mean of 3.1×10^{-5} /m through type curve analyses of 11 individual drawdown curves from observation intervals, when pumping at well PW1-4 by treating the medium to be homogeneous. Although the highest S_s value (5.88×10^{-3} /m for layer 18) of our Case 2b is slightly higher, it is still in a reasonable range (from 9.19×10^{-4} /m to 2.03×10^{-2} /m) as reported in *Batu* (*1*998) for clayey material. In later sections, all these estimates are evaluated by comparing the performances of model calibrations and validations

In terms of the reliability of estimated parameters, Tables 2 and 3 reveal that the estimated 95% confidence intervals of K and/or S_s are relatively large for some layers (e.g., layers 1* and 11 in Case 2a, and layers 1, 2, 3, 5, 17, 18 and 19 for Case 2b), which could be a result of merging layers and fixing the layer geometry during model

calibrations. PEST then forcefully estimates K and S_s for each layer and does not truncate the confidence intervals at the maximum and minimum bounds as assigned to parameters (*Doherty*, 2005).

The 95% confidence intervals are calculated on the basis of the linearity assumption that is used to derive the equations for parameter improvement in each optimization iteration, while the relationships between hydraulic head and hydraulic parameters are non-linear. Thus, a breakdown in the underpinning linearity assumption considered to calculate the confidence intervals may exaggerate the widths of the confidence intervals (*Christensen*, 1997; *Blessent et al*., 2011).

Meanwhile, unreasonably large confidence intervals imply that information or measurements provided for the optimization process are insufficient to uniquely determine these parameters (*Doherty*, 2005). Therefore, another possible reason for the unreasonably large confidence intervals in Tables 2 and 3 is that relatively few observations or even no observation data are available for layers in which the hydraulic parameters are estimated, as indicated in Fig. 2.

Large confidence intervals of K estimates were also found in a sandbox study for different geological models (*Zhao et al.*, 2016). However, confidence interval widths have been found to be reduced by providing prior information of estimated parameters (*Christensen*, 1997; *Blessent et al*., 2011; *Zhao et al*., 2016).

4.1.3 Case 3: Geostatistical inversion approach

Four scenarios are considered in the geostatistical inversion approach (Cases 3a - 3d) to test the impact of using different prior information for parameter estimations. Specifically, Case 3a starts with homogenous prior mean K and S_s fields; Cases 3b and 3c use heterogeneous prior mean K information based on geological zonations and estimated K and S_s values from model calibrations of Cases 2a and 2b, respectively. In Case 3d, the 19-layer geological model populated with K values obtained from permeameter tests is used as the prior mean K distribution.

The L_2 norm changes are plotted in the Supplementary Section as Fig. S1. We select inversion results from iteration step 110 at which the L_2 norm has stabilized, indicating the convergence of the inversion process as suggested by *Xiang et al.* (2009).

The estimated K and S_s fields for all four cases are shown in Figs. 3c - 3f and 4c - 4f, respectively. In addition, K estimates along the A-A′, B-B′, C-C′, and D-D′ cross-sections (as indicated on Fig. 1a) are extracted from the estimated K distributions of Cases 2 and 3 (Figs. S3, S4, S5, and S6 in the Supplementary Information section). Meanwhile, to facilitate the qualitative comparison of the site geology with the THT results from Cases 2 and 3, stratigraphy slices along A-A′, B-B′, C-C′, and D-D′ cross-sections (Fig. 2) are included as Figs. S3g - S6g.

In Fig. 3c (Case 3a), we find that the estimated K distribution clearly shows the double-aquifer feature in the centre of the modeling domain. More importantly, the bottom aquitard is now correctly identified as a low K zone, instead of a high K zone

as mapped in previous HT studies at the NCRS (*Berg and Illman*, 2011b, 2013, 2015; *Zhao and Illman*, 2017), when inversions relied solely on drawdown data. Unlike the previous works by *Berg and Illman* (2011b, 2013, 2015) in which they used no drawdowns observed from the low *K* layers, and the work by *Zhao and Illman* (2017) in which only limited number of steady state and quasi-steady state drawdowns were used for the SSHT analysis, we used more transient data from the low *K* zone obtained through additional pumping tests conducted at PW1-6 and PW1-7. Thus, we are feeding the inverse analysis with additional drawdown responses from the low *K* clay layers which carry non-redundant information regarding heterogeneity of the field site (*Oliver*, 1993; *Wu et al*., 2005; *Mao et al*., 2013a).

The above results indicate that, in order to obtain accurate *K* tomograms, it is necessary for future field implementations of HT to monitor drawdowns from both aquifer and aquitard layers. Otherwise, *K* estimates from permeameter (*Berg and Illman*, 2015) and/or slug/single-hole tests could be used to condition the *K* tomogram. Alternatively, complementary information other than pressure heads will have to be considered for jointly calibrating a geostatistical inverse model with pressure head data in HT analysis, such as with flowmeter tests (*Li et al.*, 2008; *Zha et al.*, 2014), seismic (*Brauchler et al.*, 2012), and/or self-potential surveys (*Soueid Ahmed et al.*, 2014). Other types of data may also be used for improving the inverse model as discussed by *Illman* (2014), but this needs to be done carefully.

The estimated *K* distributions for Cases 3b and 3c, in which the calibrated geological models are included as prior mean *K* fields in the SimSLE inversion, are

shown in Figs. 3d and 3e, respectively. We see from Case 3b (Figs. 3d and S3d - S6d) that the locations of high and low *K* zones in the domain centre are quite similar to those of Case 3a (Figs. 3c and S3c - S6c), whereas the zone shapes outside the centre are quite different. Specifically, a comparison of Cases 3a (Figs. 3c and S3c- S6c) and 3b (Figs. 3d and S3d - S6d) reveals that, (1) the low *K* zones at the top of model domain are similar in terms of locations and shapes; (2) the high and low *K* layers at the middle of model domain are recovered only for the central 15m × 15m area for Case 3a (Figs. 3c and S3c - S6c), while heterogeneity features extend to domain boundaries for Case 3b (Figs. 3d and S3d - S6d); and (3) the bottom aquitard layer is more clearly shown throughout the entire domain for Case 3b (Figs. 3d and S3d - S6d).

Similar findings are evident when comparing Case 3c (Figs. 3e and S3e - S6e) with Case 3a (Figs. 3c and S3c – S6c), in which the 19-layer geological model (Fig. 3b) is used as prior information for Case 3c instead of the 5-layer model (Fig. 3a).

As noted in Section 4.1.2, the calibration of the 5-layer geological model estimates a *K* value that is inconsistent with geological data for sand layer 11 due to the simplification of the geology and fixing the stratigraphy geometry. A comparison of estimated *K* tomograms between Cases 2a (Figs. 3a and S3a - S6a) and 3b (Figs. 3d and S3d - S6d) reveals that, the low *K* zone representing layer 11 (~ z = 10m) in Case 2a (Figs. 3a and S3a - S6a) is preserved in Case 3b (Figs. 3d and S3d - S6d) for regions near domain boundaries, where no pumping test data are available. Since SimSLE iteratively updates the prior *K* fields based on the differences between

simulated and observed pressure heads (*Xiang et al*., 2009), it is reasonable that some stratigraphic features from prior *K* distributions are preserved in the estimated *K* tomograms where the monitoring ports are not dense enough.

As shown through the comparison of Cases 2a (Figs. 3a and S3a - S6a) and 2b (Figs. 3b and S3b - S6b), the inconsistency between the estimated *K* value of the 5-layer model and geological data has to some extent been ameliorated when using the higher resolution 19-layer geological model for THT.

Meanwhile, refinement to the resolution of the *K* tomogram is evident within the central 15m × 15m area for Cases 3b (Figs. 3d and S3d - S6d) and 3c (Figs. 3e and S3e - S6e) when compared to the *K* tomograms for Cases 2a (Figs. 3a and S3a - S6a) and 2b (Figs. 3b and S3b - S6b). This is due to the sequential updating of the calibrated parameter fields with the highly parameterized geostatistical inversion approach.

The estimated *K* distribution for Case 3d is shown in Figs. 3f and S3f - S6f. We note from Figs. 3f and S3f – S6f that the estimated *K* tomogram captures the double-layer aquifer feature of the multi-aquifer-aquitard system, as seen from the other three Cases 3a, 3b and 3c. However, the very top and bottom areas in Case 3d (Figs. 3f and S3f - S6f) are now more clearly recovered as low permeable zones than those in Case 3a (Figs. 3c and S3c – S6c), Case 3b (Figs. 3d and S3d – S6d) and Case 3c (Figs. 3e and S3e – S6e).

As previously mentioned, Case 3d uses the 19-layer geological model, with each layer populated with the geometric mean of *K* values from the same layer

obtained via permeameter tests, as the prior mean K distribution for the SimSLE inversion. This case represents a particular scenario which utilized the most field (i.e., drawdowns and geological information) and laboratory (i.e., permeameter K) data. Consequently, we anticipate that Case 3d to perform the best for purposes of model validation discussed later.

The estimated K tomograms of Cases 3a through 3d collectively suggest that, when more accurate geological data are incorporated into the inversion process, the geostatistical approach yields a heterogeneity pattern more consistent with stratigraphy (Figs. S3g - S6g) in areas both near and far away from the pumping and observation wells, without incorporating additional data such as flowmeter and/or geophysical surveys. While the incorporation of geological data leads to more details in heterogeneity being recovered through SimSLE inversions, we note that the 19-layer geological model and the stratigraphy map are also interpolated from borehole logs, thus likely differ from the true stratigraphy. Therefore, the SSHT study by *Zhao et al*. (2016) and THT study by *Luo et al*. (2017) in the laboratory sandbox, where stratigraphy could be accurately mapped, were necessary in providing important insights on the usefulness of geological data for field HT analyses. Both studies found that HT analysis using accurate geological models yields results that are comparable to the highly parameterized geostatistical inverse models.

Meanwhile, the estimated S_s tomograms for Cases 3a – 3d are shown in Figs. 4c - 4f, respectively. We see from Figs. 4c - 4f that, although different prior mean K distributions are used for the four geostatistical inversion cases, the estimated S_s

tomograms, in general, do not reveal too many differences, all showing high values at the top and bottom of the domain, while low values are found in the central portion.

Physically, this is reasonable since the top and bottom areas of the modeling domain are known to consist of low K silt and clay layers, for which the S_s values are considered to be relatively higher (*Batu*, 1998) than the middle double-aquifer layer materials. In comparison to the K tomograms (Figs. 3c – 3f and Figs. S3 – S6), the S_s tomograms (Figs. 4c - 4f) are very smooth and do not reflect the aquitard layers known to be present in the middle of the double-layer aquifer zone, as indicated in the geological model (Fig. 2).

These results suggest that pumping test data alone cannot yield finer resolution heterogeneity for S_s estimates with current pumping and observation density. Additional information on S_s will be needed to obtain estimates at finer resolutions and this will be a future research topic.

4.2 Performances of Model Calibrations

Performances of different approaches are next evaluated by comparing the simulated versus observed drawdowns of eight pumping tests used for model calibrations, as plotted in Figs. 5a - 5h. A linear model is fit and included in each plot to assess the performance. Generally, the fit improves from the simple homogeneous models (Cases 1a and 1b) to the highly parameterized geostatistical inversion models (Cases 3a through 3d), with values of the coefficient of determination (R^2) increasing from 0.34 to 0.87, and the slopes of the linear model increasing from 0.19 to 0.87.

This is consistent with the results obtained by *Zhao and Illman* (2017) whom calibrated the same models, but to steady state and quasi-steady state drawdown data.

Next, model calibration performances of the different THT analysis approaches are quantitatively assessed by comparing the mean absolute error (L_1) and mean square error (L_2) norms. Those quantities are computed as:

$$L_1 = \frac{1}{n}\sum_{i=1}^{n}|\psi_i^* - \psi_i| \qquad L_2 = \frac{1}{n}\sum_{i=1}^{n}(\psi_i^* - \psi_i)^2 \tag{1}$$

where n is the total number of pressure heads used for calibration, ψ_i is the i^{th} observation head, and ψ_i^* is the corresponding simulated head. The L_1 and L_2 norms of model calibrations, as well as the corresponding ranks for all cases are provided in Fig. S2 of the Supplementary Information section. The cells of each entry in the table are color-coded to facilitate an easier comparison of different entries. In particular, we assign 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 utilize a 60-percentile value instead of the median to enhance the contrast in color. We also calculate the arithmetic mean of the L_1 and L_2 norms to rank the various models.

Examination of Fig. S2 reveals that model calibrations of the geostatistical inversion approach yield consistently better L_1 and L_2 ranking than the geological and effective parameter models. Such improvements, could be attributed to the use of highly parameterized models, which have larger degrees of freedom to adjust the model parameters to better fit the observation data.

In the previous SSHT study for a sandbox (*Illman et al*. 2015), the geological model faithfully representing the true stratigraphy was found to yield model calibration results that was similar in quality when compared to the highly parameterized geostatistical model. Here, the large difference of model calibration performances between the 19-layer geological model (Case 2b) and the geostatistical models (Cases 3a - 3d) could be attributed to: (1) the imperfect knowledge of geological zonations, since we are using stratigraphy information interpolated from discrete borehole logs; (2) calibration of a 19-layer geological model with transient data instead with steady state and quasi-steady state data as in *Illman et al*. (2015); (3) the NCRS site is highly heterogeneous with the variance (σ^2_{lnK}) estimated to be 6.50 (*Alexander et al*., 2011), while the variance (σ^2_{lnK}) of the sandbox investigated by *Illman et al*. (2015) is much lower, estimated to be between 0.38 and 1.32 depending on the approach used for characterization (*Berg and Illman*, 2011a).

Comparisons among different geostatistical inverse models (in Figs. 5 and S2) reveal that, when the geostatistical inversion approach starts from the geologically distributed K fields (Cases 3b, 3c and 3d) instead of the uniform K value (Case 3a), the calibration performances are generally improved as indicated by the fitting parameters (R^2, L_1 and L_2) of the linear model.

Previously, *Berg and Illman* (2015) performed a comparative study of different traditional methods with THT in terms of characterizing the heterogeneity of the aquifer-aquitard system. Specifically, in order to correctly identify the low K clayey zones near the bottom of the modelling domain, they conditioned the geostatistical

model to the permeameter *K* data by fixing the *K* values of computational elements along five boreholes (PW1, CMT1 to CMT4) during the inversion process. Although the estimated *K* distributions were found to be consistent with known geology, the THT inversions also yielded slight deterioration in model calibration when permeameter *K* data were used to condition the model. Unlike the approach adopted by *Berg and Illman* (2015), geological information was used as the starting *K* distributions of the geostatistical inversions in this study and in *Zhao and Illman* (2017). Thus, the improved model calibration of Cases 3b, 3c and 3d compared to Case 3a shown in Fig. 5, is a result of both including reliable geological data and enabling the pumping tests data to freely update the prior mean values.

Here, we need to clarify that, due to the availability of a large number of borehole logs and laboratory measurement data at the NCRS, the geological model based on the interpolation between wells may be more reliable. In contrast, at other sites, the well network may be sparse and the geological data could contain large errors (e.g., boundary locations, zone structures, misidentification of layers, etc.). Under such circumstances, a general framework proposed by *Zha et al*. (2017) that allows the inclusion of site-specific geologic features/hydraulic properties as well as their associated uncertainties, could be adopted to further improve HT results. Their framework adopts a nested covariance function to conceptualize the site heterogeneity, which requires site-specific geological information at both small and large scales to construct the prior mean and covariance. Unlike the approach proposed by *Zha et al*.

(2017), the approach that we utilize in this study uses kriging of borehole data to construct geological models and is very practical.

4.3 Comparison of Estimated *K* with Permeameter Test *K*

In order to further examine the ability of geological and geostatistical inverse models to capture the heterogeneity of the NCRS, the estimated *K* values are extracted from the THT tomograms and compared with permeameter test *K* values along all PW and CMT wells (Figs. 6 and S7). Meanwhile, the *K* estimates from SSHT analysis of seven pumping tests starting with a uniform *K* value of $K = 8.0 \times 10^{-6}$ m/s by *Zhao and Illman* (2017) are jointly plotted as red lines in Figs. 6 and S7 to provide direct comparisons with the THT estimates.

As seen in Figs. 6 and S7, calibrations of the 5-layer and 19-layer geological models to eight pumping tests yield *K* estimates (blue lines) that generally follow the vertical variations of permeameter test *K* values (black dashed lines with dots). However, the intralayer heterogeneity in *K* values of each unit is not captured by both geological models due to the assumption of constant *K* values in each layer. Still, the *K* profile estimated by the 19-layer geological model reveals more variations than the simplified 5-layer geological model. We also see that the results of eight-test THT analysis (blue lines) are quite similar to those of the seven-test SSHT steady state analysis (red lines), although there are minor differences. This finding suggests that both geological models with low degrees of parameterizations for this highly heterogeneous site may be justified (*Schöniger et al.*, 2015) given the availability of multiple pumping test data to conduct the inverse modeling at the NCRS.

On the other hand, the calibration of the highly parameterized geostatistical inverse model starting from the uniform mean *K* field (Case 3a) captures the overall changes in the *K* profile from the top to the bottom of all PW and CMT wells. The most striking improvement is that the transient inversion results (blue lines) fit the permeameter *K* measurements quite well at depths of 0 - 5 m and 10 - 15 m above the bottom of the modeling domain, whereas relatively high *K* values (red lines) are obtained by *Zhao and Illman* (2017) when only using steady state data. This result indicates that the inclusion of additional pumping tests conducted at aquitard layers (i.e., PW1-1, PW1-6, and PW1-7) and more transient drawdown data into the inversion has led to the correct identification of aquitard zones and the improved characterization of intralayer *K* heterogeneity.

In Cases 3b, 3c and 3d, the geologically distributed prior *K* mean fields are used as starting values for the geostatistical inversion approach. Compared to the results of Case 3a in Figs. 6 and S7, Cases 3b, 3c and 3d yield *K* profiles that are more consistent with the permeameter test *K* values, in terms of the locations and thicknesses of aquifer and aquitard layers. That is, the boundaries between aquifer and aquitard layers are better delineated in Cases 3b, 3c and 3d than those in Case 3a.

Meanwhile, unlike the significant differences between the results of steady state and transient analyses in Case 3a, the blue and red lines representing the vertical *K* variations in Cases 3b, 3c and 3d in general show satisfactory matches, especially for Cases 3c and 3d in which 19-layer geological models are used as prior *K* distributions.

These results collectively suggest that the inverse modeling of transient pumping test data can yield reliable heterogeneous distributions of *K* for both aquifer and aquitards including its intralayer heterogeneity even at a highly heterogeneous site such as at the NCRS. Joint inverse modeling of transient drawdown data from aquifer and aquitard layers is necessary in accurately characterizing the site with a highly parameterized geostatistical approach. Moreover, using reliable geological models as initial distributions are helpful for the geostatistical inversion approach of THT in improving the correspondence of estimated *K* for both the aquifer and aquitard layers to those from permeameter tests.

4.4 Model Validation Results

We next evaluate the performances of different models in their abilities of predicting pumping tests not used during the calibration process. As previously noted, four pumping tests (PW1-3, PW1-5, PW5-4, and PW5-5) are selected for model validation purposes. The drawdown predictions of different models and various tests are plotted as Figs. 7 (PW1-3), S8 (PW1-5), S9 (PW5-4), and S10 (PW5-5). Meanwhile, the simulated drawdowns are extracted at selected times and compared with corresponding observed drawdowns to provide more quantitative comparisons (Figs. 8 and S11).

In Figs. 7, S8, S9 and S10, the drawdown predictions of the homogeneous/isotropic model (Case 1a: solid gray lines), the homogeneous/anisotropic model (Case 1b: dashed gray lines), the 5-layer geological model (Case 2a: dash-dotted gray lines), the 19-layer geological model (Case 2b:

dotted black lines) and different geostatistical models (Cases 3a - 3d: lines in red color) are compared to the drawdowns observed (solid blue lines with open circles) from different locations of the aquifer-aquitard system. Through visual examinations of the match quality between different simulated drawdowns to observed curves in Fig. 7 for the pumping test at PW1-3, we find that geostatistical models of Case 3 yield drawdown predictions that are close to each other and in general, capture the blue lines observed from different observation intervals. In contrast, obvious poor matches are found between drawdown predictions of the homogeneous models (solid and dashed gray lines) and the observations (blue lines), either at early time or late time of the drawdown curves. On the other hand, the simplified 5-layer geological model (Case 2a) consistently over-predicts the drawdowns for observation ports (e.g., CMT1-1, CMT2-1, and PW5-1) located within the top layer of the aquifer-aquitard system, and also yield poor matches to the blue curves in the rest of observation intervals as the homogeneous models (Cases 1a and 1b). When the complex 19-layer geological model (Case 2b) is used, the match qualities improve for drawdown predictions over the 5-layer model (Case 2a). At several observation intervals, the matches are comparable to those of the highly parameterized geostatistical models (Cases 3a - 3d).

In Fig. 8, the observed drawdowns are compared to the simulated values selected from different monitoring intervals and times indicated by blue circles in Figs. 7, S8, S9, and S10. The linear fits show that the homogeneous/isotropic model (Case 1a) yields biased predictions (Fig. 8a) and underestimates the drawdowns of four

pumping tests, with a coefficient of determination (R^2) value of 0.23 and a very shallow slope for the linear model fit with a value of 0.08. The prediction results are improved slightly when using either the homogeneous/anisotropic model (Case 1b) or the simplified 5-layer geological model (Case 2a), as shown in Figs. 8b and 8c. In addition, performances of the 19-layer geological model (Case 2b) show obvious improvements over the above three models (Cases 1a, 1b and 2a), with the R^2 and the slope of the linear model fit increasing to 0.65 and 1.15, respectively.

It is worth noting that the 19-layer geological model (Case 2b) performed closely to the geostatistical inversion approach (Case 3a) starting with $K = 8.0 \times 10^{-6}$ m/s as a prior mean. On the other hand, compared to the effective parameter (Cases 1a and 1b) and the geological modeling (Cases 2a and 2b) approaches, models of the highly parameterized geostatistical inversion approach using geologically distributed K prior means (Cases 3b – 3d) yield consistently improved R^2 and L_1, while the L_2 ranking behaves differently.

The lowest L_2 norm is reached by the geostatistical inversion model using the calibrated 19-layer geological model as prior K distribution (Case 3c). Since the L_2 norm magnifies large discrepancies between the simulated and the observed drawdowns compared to the L_1 norm, the high L_2 norm values in Cases 3a, 3b and 3c mainly represent the impact of poor fittings of drawdowns at several observation intervals such as at CMT2-2 and PW3-3 in Fig. 7. Such an inconsistent L_2 ranking among Cases 3a - 3d indicate that, while the K (Fig. 3) and S_s (Fig. 4) tomograms show similar overall patterns, the fine-scale differences in these K and S_s tomograms

could lead to drastically different drawdown predictions at some observation intervals. Thus, the geostatistical inversion approach that can correctly capture both interlayer and intralayer heterogeneity should be more favored for future HT applications.

5. Summary and Conclusions

In this study, we investigated whether transient drawdown data collected in both the aquifer and aquitard units can be used in transient hydraulic tomography (THT) analysis to reliably map the interlayer and intralayer heterogeneity of these units, at a field site underlain by a highly heterogeneous multi-aquifer-aquitard system. In addition, we examined the value of geological data for THT analysis by constructing geological zonation models based on available stratigraphy information and using them as prior distributions for the geostatistical inversion approach. We selected 12 pumping tests for model calibration and validation. Eight tests containing hydraulic response information observed in aquifer and aquitard layers are jointly used to calibrate groundwater models of different complexities and parameterization schemes, including: (1) two effective parameter models (isotropic/anisotropic); (2) two geological zonation models of varying numbers of layers and (3) four highly parameterized geostatistical inverse models that consider different prior K means. Our study resulted in the following major findings and conclusions:

1. Despite the large number of data used for model calibration, the effective parameter models calibrated to eight pumping tests yielded biased calibration and validation results. In contrast, THT analyses based on geological models and highly parameterized geostatistical models led to much improved model calibration and

drawdown prediction performances. This suggests that it may be difficult to represent site heterogeneity with an effective parameter model at the scale of site investigation conducted at the NCRS.

2. Although the layer boundaries were fixed, geological models calibrated to eight pumping test data yielded *K* estimates that were close to the general patterns of vertical *K* distributions from permeameter tests. Meanwhile, the calibration of the geostatistical inversion model to drawdown data with initially uniform hydraulic parameter fields correctly recovered both the low and high *K* zones within the well field, although smoothed results were obtained for areas where no drawdown data were available.

3. Instead of using an effective *K* estimate, calibrated groundwater models based on geological zonation used as an initial guess was very helpful for the geostatistical inversion approach in improving the correspondence of estimated *K* to those from permeameter tests, and in preserving geological features/connectivity in *K* heterogeneity patterns. Different from the previous hydraulic tomography study by *Zhao and Illman* (2017) who relied on steady state inversions, the joint inverse modeling of transient drawdown data from aquifer and aquitard layers yielded reliable heterogeneous distributions of *K* and S_s for both aquifer and aquitards even at a highly heterogeneous site.

4. We conclude that the inclusion of reliable geological data is useful for THT analysis to better image the full heterogeneity patterns of complex groundwater systems. In addition, sufficiently long records of pumping and observation data should

be collected not only from aquifer layers, but also from aquitard layers for future HT studies when relying only on pumping test data.

5. Collection of long drawdown records will require the judicious selection of data that should be fed into inverse models to avoid information overload, and inclusion of redundant as well as erroneous data. Therefore, it is necessary to systematically examine the quality of data, the information content of observed heads from both high and low *K* zones, as well as the value of other types of complementary data. Data-worth estimation methods such as the Preposterior Data Impact Assessor (PreDIA) (*Leube et al.*, 2012) and the linear predictive uncertainty quantification (PREDUNC) utility (*Wöhling et al*., 2016) may be applied to evaluate monitoring strategies and to identify potentially redundant information, both of which can potentially improve the quality of HT analyses.

Acknowledgements: This research was supported by the Environmental Security and Technology Certification Program (ESTCP) under grant ER201212 to Walter A. Illman. The first author acknowledges the support of ‘‘China Scholarship Council”. We thank the project principal investigator C. M. Mok and co-investigator T.-C. J. Yeh for fruitful discussions on the designs of some pumping/injection tests used in this study. Additional support for the project was provided to Walter A. Illman by the Discovery and Collaborative Research and Development Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC), as well as grants from the Ontario Research Foundation (ORF) and Canada Foundation for Innovation

(CFI). We thank the Associate Editor and three anonymous reviewers for their helpful comments in improving the manuscript

References

Alexander, M., Berg, S.J., Illman, W.A., 2011. Field Study of Hydrogeologic Characterization Methods in a Heterogeneous Aquifer. Ground Water 49, 365–382. https://doi.org/10.1111/j.1745-6584.2010.00729.x

ARANZ Geo. Limited., 2015. Leapfrog Hydro 2.2.3. 3D Geological Modelling Software.

Batu, V., 1998. Aquifer hydraulics: a comprehensive guide to hydrogeologic data analysis. John Wiley & Sons.

Berg, S.J., Hsieh, P.A., Illman, W.A., 2011. Estimating Hydraulic Parameters When Poroelastic Effects Are Significant. Ground Water 49, 815–829. https://doi.org/10.1111/j.1745-6584.2010.00781.x

Berg, S.J., Illman, W.A., 2015. Comparison of Hydraulic Tomography with Traditional Methods at a Highly Heterogeneous Site. Groundwater 53, 71–89. https://doi.org/10.1111/gwat.12159

Berg, S.J., Illman, W.A., 2013. Field study of subsurface heterogeneity with steady-state hydraulic tomography. GroundWater 51, 29–40. https://doi.org/10.1111/j.1745-6584.2012.00914.x

Berg, S.J., Illman, W.A., 2011a. Capturing aquifer heterogeneity: Comparison of approaches through controlled sandbox experiments. Water Resour. Res. 47, 1–17. https://doi.org/10.1029/2011WR010429

Berg, S.J., Illman, W.A., 2011b. Three-dimensional transient hydraulic tomography in a highly heterogeneous glaciofluvial aquifer-aquitard system. Water Resour. Res. 47. https://doi.org/10.1029/2011WR010616

Berg, S.J., Illman, W.A., Mok, C.M.W., 2015. Joint Estimation of Hydraulic and Poroelastic Parameters from a Pumping Test. Groundwater 53, 759–770. https://doi.org/10.1111/gwat.12271

Blessent, D., Therrien, R., Lemieux, J.M., 2011. Inverse modeling of hydraulic tests in fractured crystalline rock based on a transition probability geostatistical approach. Water Resour. Res. 47, 1–19. https://doi.org/10.1029/2011WR011037

Bohling, G.C., Butler, J.J., Zhan, X., Knoll, M.D., 2007. A field assessment of the value of steady shape hydraulic tomography for characterization of aquifer

heterogeneities. Water Resour. Res. 43, 1–23. https://doi.org/10.1029/2006WR004932

Bohling, G.C., Zhan, X., Butler, J.J., Zheng, L., 2002. Steady shape analysis of tomographic pumping tests for characterization of aquifer heterogeneities. Water Resour. Res. 38, 60-1-60–15. https://doi.org/10.1029/2001WR001176

Brauchler, R., Doetsch, J., Dietrich, P., Sauter, M., 2012. Derivation of site-specific relationships between hydraulic parameters and p-wave velocities based on hydraulic and seismic tomography. Water Resour. Res. 48, 1–14. https://doi.org/10.1029/2011WR010868

Brauchler, R., Hu, R., Dietrich, P., Sauter, M., 2011. A field assessment of high-resolution aquifer characterization based on hydraulic travel time and hydraulic attenuation tomography. Water Resour. Res. 47, 1–12. https://doi.org/10.1029/2010WR009635

Brauchler, R., Liedl, R., Dietrich, P., 2003. A travel time based hydraulic tomographic approach. Water Resour. Res. 39, 1370. https://doi.org/10.1029/2003WR002262

Bredehoeft, J.D., Neuzil, C.E., Milly, P.C., 1983. Regional flow in the Dakota aquifer; A study of the role of confining layers. USGPO,.

Cardiff, M., Bakhos, T., Kitanidis, P.K., Barrash, W., 2013. Aquifer heterogeneity characterization with oscillatory pumping: Sensitivity analysis and imaging potential. Water Resour. Res. 49, 5395–5410. https://doi.org/10.1002/wrcr.20356

Cardiff, M., Barrash, W., 2011. 3-D transient hydraulic tomography in unconfined aquifers with fast drainage response. Water Resour. Res. 47. https://doi.org/10.1029/2010WR010367

Cardiff, M., Barrash, W., Kitanidis, P.K., 2013. Hydraulic conductivity imaging from 3-D transient hydraulic tomography at several pumping/observation densities. Water Resour. Res. 49, 7311–7326. https://doi.org/10.1002/wrcr.20519

Cardiff, M., Barrash, W., Kitanidis, P.K., 2012. A field proof-of-concept of aquifer imaging using 3-D transient hydraulic tomography with modular, temporarily-emplaced equipment. Water Resour. Res. 48. https://doi.org/10.1029/2011WR011704

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

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., Slooten, L.J., 2005. Inverse problem in hydrogeology. Hydrogeol. J. 13, 206–222. https://doi.org/10.1007/s10040-004-0404-7

Christensen, S., 1997. On the strategy of estimating regional-scale transmissivity fields. Ground Water. https://doi.org/10.1111/j.1745-6584.1997.tb00068.x

Cleveland, T.G., Bravo, R., Rogers, J.R., 1992. Storage Coefficients and Vertical Hydraulic Conductivities in Aquitards Using Extensometer and Hydrograph Data. Groundwater 30, 701–708. https://doi.org/10.1111/j.1745-6584.1992.tb01556.x

Davis, R.W., 1972. Use of naturally occurring phenomena to study hydraulic diffusivities of aquitards. Water Resour. Res. 8, 500–507. https://doi.org/10.1029/WR008i002p00500

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

Fogg, G.E., Zhang, Y., 2016. Debates—Stochastic subsurface hydrology from theory to practice: A geologic perspective. Water Resour. Res. 52, 9235–9245. https://doi.org/10.1002/2016WR019699

Galloway, D.L., Burbey, T.J., 2011. Review: Regional land subsidence accompanying groundwater extraction. Hydrogeol. J. 19, 1459–1486. https://doi.org/10.1007/s10040-011-0775-5

Ghorbanidehno, H., Kokkinaki, A., Kitanidis, P.K., Darve, E., 2017. Optimal estimation and scheduling in aquifer management using the Rapid Feedback Control Method. Adv. Water Resour. 110, 310–318. https://doi.org/10.1016/j.advwatres.2017.10.011

Ghorbanidehno, H., Kokkinaki, A., Li, J.Y., Darve, E., Kitanidis, P.K., 2015. Real-time data assimilation for large-scale systems: The spectral Kalman filter. Adv. Water Resour. 86, 260–272. https://doi.org/10.1016/j.advwatres.2015.07.017

Hendry, M.J., Wassenaar, L.I., 2000. Controls on the distribution of major ions in pore waters of a thick surficial aquitard. Water Resour. Res. 36, 503–513. https://doi.org/10.1029/1999WR900310

Hochstetler, D.L., Barrash, W., Leven, C., Cardiff, M., Chidichimo, F., Kitanidis, P.K., 2016. Hydraulic Tomography: Continuity and Discontinuity of High-K and Low-K Zones. Ground Water 54, 171–85. https://doi.org/10.1111/gwat.12344

Huang, S.Y., Wen, J.C., Yeh, T.C.J., Lu, W., Juan, H.L., Tseng, C.M., Lee, J.H., Chang, K.C., 2011. Robustness of joint interpretation of sequential pumping tests: Numerical and field experiments. Water Resour. Res. 47, 1–18. https://doi.org/10.1029/2011WR010698

Illman, W.A., 2014. Hydraulic tomography offers improved imaging of heterogeneity in fractured rocks. Groundwater 52, 659–684. https://doi.org/10.1111/gwat.12119

Illman, W.A., Berg, S.J., Yeh, T.C.J., 2012. Comparison of Approaches for Predicting Solute Transport: Sandbox Experiments. Ground Water 50, 421–431. https://doi.org/10.1111/j.1745-6584.2011.00859.x

Illman, W.A., Berg, S.J., Zhao, Z., 2015. Should hydraulic tomography data be interpreted using geostatistical inverse modeling? A laboratory sandbox investigation. Water Resour. Res. 51, 3219–3237. https://doi.org/10.1002/2014WR016552

Illman, W.A., Liu, X., Takeuchi, S., Yeh, T.C.J., Ando, K., Saegusa, H., 2009. Hydraulic tomography in fractured granite: Mizunami Underground Research site, Japan. Water Resour. Res. 45, 1–18. https://doi.org/10.1029/2007WR006715

Illman, W.A., Zhu, J., Craig, A.J., Yin, D., 2010. Comparison of aquifer characterization approaches through steady state groundwater model validation: A controlled laboratory sandbox study. Water Resour. Res. 46, 1–18. https://doi.org/10.1029/2009WR007745

Illman, W. a, Liu, X., Craig, A., 2008. Evaluation of transient hydraulic tomography and common hydraulic characterization approaches through laboratory sandbox experiments. J. Environ. Eng. Manag. 18, 249–256.

Johnson, R.L., Cherry, J.A., Pankow, J.F., 1989. Diffusive contaminant transport in natural clay: a field example and implications for clay-lined waste disposal sites. Environ. Sci. Technol. 23, 340–349. https://doi.org/10.1021/es00180a012

Karrow, P.F., 1979. Geology of the University of Waterloo Campus. Waterloo, Ontario, Canada.

Keller, C.K., Van Der Kamp, G., Cherry, J.A., 1989. A multiscale study of the permeability of a thick clayey till. Water Resour. Res. 25, 2299–2317. https://doi.org/10.1029/WR025i011p02299

Kitanidis, P.K., 2015. Compressed state Kalman filter for large systems. Adv. Water Resour. 76, 120–126. https://doi.org/10.1016/j.advwatres.2014.12.010

Kitanidis, P.K., 1995. Quasi-linear geostatistical theory for inversing. Water Resour. Res. https://doi.org/10.1029/95WR01945

Konikow, L.F., Neuzil, C.E., 2007. A method to estimate groundwater depletion from confining layers. Water Resour. Res. 43, n/a-n/a. https://doi.org/10.1029/2006WR005597

Leube, P.C., Geiges, A., Nowak, W., 2012. Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design 48, 1–16. https://doi.org/10.1029/2010WR010137

Li, J.Y., Kokkinaki, A., Ghorbanidehno, H., Darve, E.F., Kitanidis, P.K., 2015. The

compressed state Kalman filter for nonlinear state estimation: Application to large-scale reservoir monitoring. Water Resour. Res. 51, 9942–9963. https://doi.org/10.1002/2015WR017203

Li, L., Zhou, H., Gómez-Hernández, J.J., Hendricks Franssen, H.-J., 2012. Jointly mapping hydraulic conductivity and porosity by assimilating concentration data via ensemble Kalman filter. J. Hydrol. 428–429, 152–169. https://doi.org/10.1016/j.jhydrol.2012.01.037

Li, W., Englert, A., Cirpka, O.A., Vereecken, H., 2008. Three-dimensional geostatistical inversion of flowmeter and pumping test data. Ground Water 46, 193–201. https://doi.org/10.1111/j.1745-6584.2007.00419.x

Liu, S., Yeh, T.C.J., Gardiner, R., 2002. Effectiveness of hydraulic tomography: Sandbox experiments. Water Resour. Res. 38, 2–10. https://doi.org/10.1029/2001WR000338

Liu, X., Illman, W.A., Craig, A.J., Zhu, J., Yeh, T.C.J., 2007. Laboratory sandbox validation of transient hydraulic tomography. Water Resour. Res. 43, 1–13. https://doi.org/10.1029/2006WR005144

Liu, X., Kitanidis, P.K., 2011. Large-scale inverse modeling with an application in hydraulic tomography. Water Resour. Res. 47, 1–9. https://doi.org/10.1029/2010WR009144

Luo, N., Zhao, Z., Illman, W.A., Berg, S.J., 2017. Comparative study of transient hydraulic tomography with varying parameterizations and zonations: Laboratory sandbox investigation. J. Hydrol. 554, 758–779. https://doi.org/https://doi.org/10.1016/j.jhydrol.2017.09.045

Mao, D., Yeh, T.C.J., Wan, L., Lee, C.H., Hsu, K.C., Wen, J.C., Lu, W., 2013a. Cross-correlation analysis and information content of observed heads during pumping in unconfined aquifers. Water Resour. Res. 49, 713–731. https://doi.org/10.1002/wrcr.20066

Mao, D., Yeh, T.C.J., Wan, L., Wen, J.C., Lu, W., Lee, C.H., Hsu, K.C., 2013b. Joint interpretation of sequential pumping tests in unconfined aquifers. Water Resour. Res. 49, 1782–1796. https://doi.org/10.1002/wrcr.20129

Martin, P.J., Frind, E.G., 1998. Modeling a Complex Multi-Aquifer System: The Waterloo Moraine. Ground Water 36, 679–690. https://doi.org/10.1111/j.1745-6584.1998.tb02843.x

Montazer, P., Wilson, W.E., 1984. Conceptual hydrologic model of flow in the unsaturated zone, Yucca Mountain, Nevada. U.S. Geol. Surv. Water Resour. Invest. Rep.84-4355.

Neuman, S.P., Witherspoon, P.A., 1972. Field determination of the hydraulic

properties of leaky multiple aquifer systems. Water Resour. Res. 8, 1284–1298. https://doi.org/10.1029/WR008i005p01284

Neuzil, C.E., 1994. How permeable are clays and shales? Water Resour. Res. 30, 145–150. https://doi.org/10.1029/93WR02930

Neuzil, C.E., 1986. Groundwater Flow in Low-Permeability Environments. Water Resour. Res. 22, 1163–1195. https://doi.org/10.1029/WR022i008p01163

Nowak, W., 2009. Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator. Water Resour. Res. 45, 1–17. https://doi.org/10.1029/2008WR007328

Oliver, D.S., 1993. The influence of nonuniform transmissivity and storativity on drawdown. Water Resour. Res. 29, 169–178. https://doi.org/10.1029/92WR02061

Paradis, D., Gloaguen, E., Lefebvre, R., Giroux, B., 2016. A field proof-of-concept of tomographic slug tests in an anisotropic littoral aquifer. J. Hydrol. 536, 61–73. http://dx.doi.org/10.1016/j.jhydrol.2016.02.041.

Panzeri, M., Riva, M., Guadagnini, A., Neuman, S.P., 2013. Data assimilation and parameter estimation via ensemble Kalman filter coupled with stochastic moment equations of transient groundwater flow. Water Resour. Res. 49, 1334–1344. https://doi.org/10.1002/wrcr.20113

Pool, M., Carrera, J., Alcolea, A., Bocanegra, E.M., 2015. A comparison of deterministic and stochastic approaches for regional scale inverse modeling on the Mar del Plata aquifer. J. Hydrol. 531, 214–229. https://doi.org/10.1016/j.jhydrol.2015.09.064

Schöniger, A., Illman, W.A., Wöhling, T., Nowak, W., 2015. Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. J. Hydrol. 531, 96–110. https://doi.org/10.1016/j.jhydrol.2015.07.047

Schöniger, A., Nowak, W., Hendricks Franssen, H.J., 2012. Parameter estimation by ensemble Kalman filters with transformed data: Approach and application to hydraulic tomography. Water Resour. Res. 48, 1–18. https://doi.org/10.1029/2011WR010462

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

Soueid Ahmed, A., Jardani, A., Revil, A., Dupont, J.P., 2014. Hydraulic conductivity field characterization from the joint inversion of hydraulic heads and self-potential data. Water Resour. Res. 50, 3502–3522.

https://doi.org/10.1002/2013WR014645

Soueid Ahmed, A., Zhou, J., Jardani, A., Revil, A., Dupont, J.P., 2015. Image-guided inversion in steady-state hydraulic tomography. Adv. Water Resour. 82, 83–97. https://doi.org/10.1016/j.advwatres.2015.04.001

Therrien, R., Mclaren, R.G., Sudicky, E.A., Panday, S.M., 2005. HydroGeoSphere: A Three-dimensional Numerical Model Describing Fully-integrated Subsurface and Surface Flow and Solute Transport.

Tso, C.-H.M., Yeh, T.-C.J., Zha, Y., Wen, J.-C., 2016. The relative importance of head, flux, and prior information in hydraulic tomography analysis. Water Resour. Res. 52, 3–20. https://doi.org/10.1002/2015WR017191

Van der Kamp, G., 2001. Methods for determining the in situ hydraulic conductivity of shallow aquitards - An overview. Hydrogeol. J. https://doi.org/10.1007/s100400000118

van der Kamp, G., Maathuis, H., 1985. Excess hydraulic head in aquitards under solid waste emplacements. Hydrogeol. rocks low permeability 17, 118.

Verruijt, A., 1969. Elastic storage of aquifers. Flow through porous media 331–376.

Wen, J.C., Wu, C.M., Yeh, T.C.J., Tseng, C.M., 2010. Estimation of effective aquifer hydraulic properties from an aquifer test with multi-well observations. Hydrogeol. J. 18, 1143–1155. https://doi.org/10.1007/s10040-010-0577-1

Wen, X.-H., Chen, W.H., 2006. Real-Time Reservoir Model Updating Using Ensemble Kalman Filter With Confirming Option. SPE J. 11, 431–442. https://doi.org/10.2118/92991-PA

Wöhling, T., Geiges, A., Nowak, W., 2016. Optimal Design of Multitype Groundwater Monitoring Networks Using Easily Accessible Tools. Groundwater 54, 861–870. https://doi.org/10.1111/gwat.12430

Wu, C.M., Yeh, T.C.J., Zhu, J., Tim, H.L., Hsu, N.S., Chen, C.H., Sancho, A.F., 2005. Traditional analysis of aquifer tests: Comparing apples to oranges? Water Resour. Res. 41, 1–12. https://doi.org/10.1029/2004WR003717

Xiang, J., Yeh, T.C.J., Lee, C.H., Hsu, K.C., Wen, J.C., 2009. A simultaneous successive linear estimator and a guide for hydraulic tomography analysis. Water Resour. Res. 45, 1–14. https://doi.org/10.1029/2008WR007180

Yeh, T.C.J., Jin, M., Hanna, S., 1996. An Iterative Stochastic Inverse Method: Conditional Effective Transmissivity and Hydraulic Head Fields. Water Resour. Res. 32, 85–92. https://doi.org/10.1029/95WR02869

Yeh, T.C.J., Liu, S., 2000. Hydraulic tomography: Development of a new aquifer test method. Water Resour. Res. 36, 2095. https://doi.org/10.1029/2000WR900114

Yeh, T.C.J., Mao, D. qiang, Zha, Y. yuan, Wen, J. chau, Wan, L., Hsu, K. chin, Lee, C. haw, 2015. Uniqueness, scale, and resolution issues in groundwater model parameter identification. Water Sci. Eng. 8, 175–194. https://doi.org/10.1016/j.wse.2015.08.002

Yeh, T.C.J., Šimůnek, J., 2002. Stochastic Fusion of Information for Characterizing and Monitoring the Vadose Zone. Vadose Zo. J. 1, 207. https://doi.org/10.2136/vzj2002.0207

Zha, Y., Yeh, T.-C.J., Illman, W.A., Onoe, H., Man, C., Mok, W., Wen, J.-C., Huang, S.-Y., Wang, W., 2017. Incorporating geologic information into hydraulic tomography: A general framework based on geostatistical approach. Water Resour. Res. n/a-n/a. https://doi.org/10.1002/2016WR019185

Zha, Y., Yeh, T.C.J., Mao, D., Yang, J., Lu, W., 2014. Usefulness of flux measurements during hydraulic tomographic survey for mapping hydraulic conductivity distribution in a fractured medium. Adv. Water Resour. 71, 162–176. https://doi.org/10.1016/j.advwatres.2014.06.008

Zhao, Z., Illman, W.A., 2017. On the importance of geological data for three-dimensional steady-state hydraulic tomography analysis at a highly heterogeneous aquifer-aquitard system. J. Hydrol. 544, 640–657. https://doi.org/http://dx.doi.org/10.1016/j.jhydrol.2016.12.004

Zhao, Z., Illman, W.A., Berg, S.J., 2016. On the importance of geological data for hydraulic tomography analysis: Laboratory sandbox study. J. Hydrol. https://doi.org/http://dx.doi.org/10.1016/j.jhydrol.2016.08.061

Zhao, Z., Illman, W.A., Yeh, T.C.J., Berg, S.J., Mao, D., 2015. Validation of hydraulic tomography in an unconfined aquifer: A controlled sandbox study. Water Resour. Res. 51, 4137–4155. https://doi.org/10.1002/2015WR016910

Zhou, H., Gómez-Hernández, J.J., Li, L., 2014. Inverse methods in hydrogeology: Evolution and recent trends. Adv. Water Resour. 63, 22–37. https://doi.org/10.1016/j.advwatres.2013.10.014

Zhou, Y., Lim, D., Cupola, F., Cardiff, M., 2016. Aquifer imaging with pressure waves-Evaluation of low-impact characterization through sandbox experiments. Water Resour. Res. n/a-n/a. https://doi.org/10.1002/2015WR017751

Zhu, J., Yeh, T.C.J., 2006. Analysis of hydraulic tomography using temporal moments of drawdown recovery data. Water Resour. Res. 42, 1–11. https://doi.org/10.1029/2005WR004309

Zhu, J., Yeh, T.C.J., 2005. Characterization of aquifer heterogeneity using transient hydraulic tomography. Water Resour. Res. 41, 1–10. https://doi.org/10.1029/2004WR003790

Zhuang, C., Zhou, Z., Zhan, H., Wang, G., 2015. A new type curve method for estimating aquitard hydraulic parameters in a multi-layered aquifer system. J. Hydrol. 527, 212–220. https://doi.org/https://doi.org/10.1016/j.jhydrol.2015.04.062

Zhuang, C., Zhou, Z., Illman, W.A., 2017a. A Joint Analytic Method for Estimating Aquitard Hydraulic Parameters. Groundwater n/a-n/a. https://doi.org/10.1111/gwat.12494

Zhuang, C., Zhou, Z., Illman, W.A., Guo, Q., Wang, J., 2017b. Estimating hydraulic parameters of a heterogeneous aquitard using long-term multi-extensometer and groundwater level data. Hydrogeol. J. https://doi.org/10.1007/s10040-017-1596-y

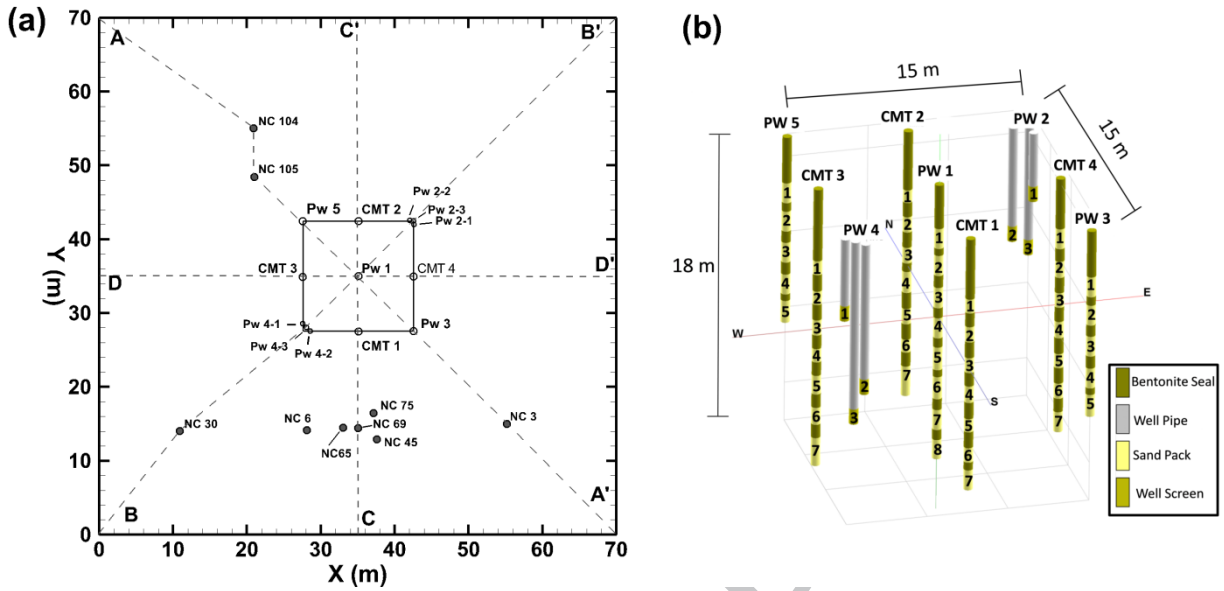


Fig. 1. (a) Plan view showing well locations used in this study at the North Campus Research Site (NCRS) situated on the University of Waterloo (UW) campus. Solid circles indicate the locations where only geological data are available. Dashed lines indicate four geological cross sections: A-A′, B-B′, C-C′ and D-D′ provided in Fig. 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).

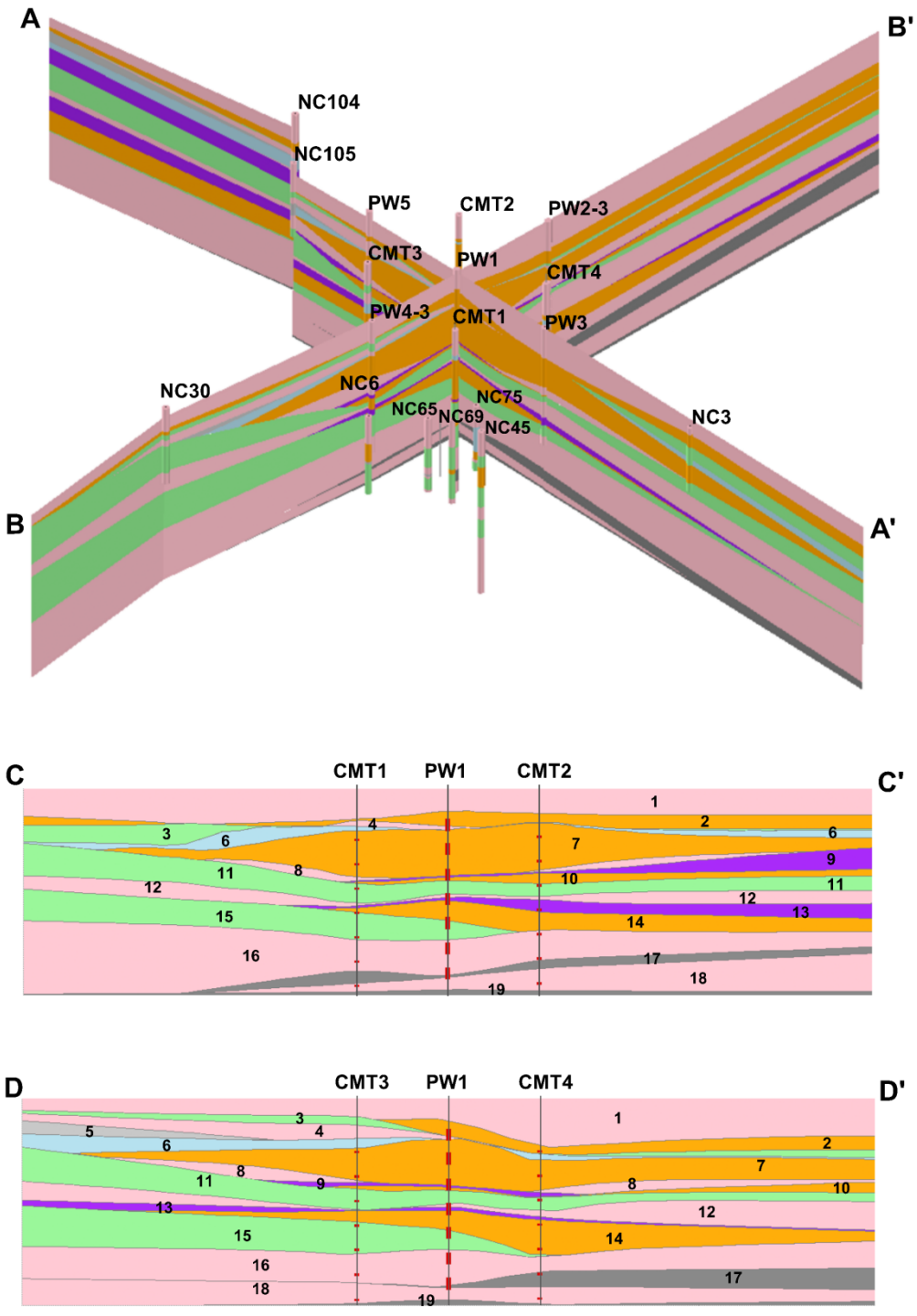


Fig. 2. Cross sections of the geological model: A-A′, B-B′, C-C′, and D-D′, at the NCRS. Numbers in cross section C-C′ and D-D′ indicate the 19 layers of different materials: Clay (1, 4, 8, 12, 16, 18); Silt and Clay (17, 19); Silt (2, 7, 10, 14); Sandy Silt (6, 9, 13); Sand and Silt (5); Sand (3, 11); Sand and Gravel (15). Screened locations are shown on wells depicted in cross sections C-C′ and D-D′ (from *Zhao and Illman*, 2017).

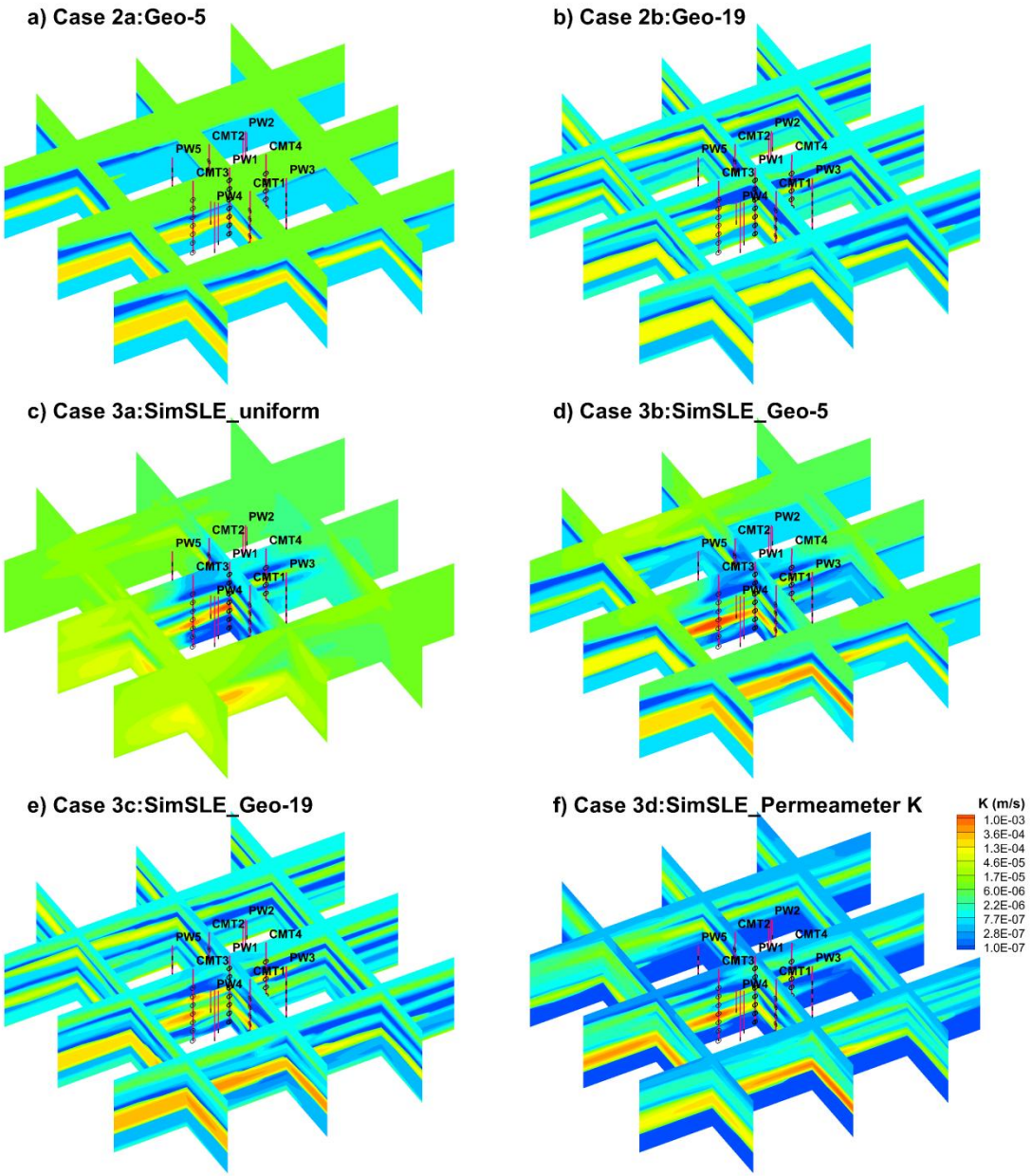


Fig. 3. Estimated *K* fields from the inversion of eight pumping tests for: (a) Case 2a: the 5-layer geological model; (b) Case 2b: the 19-layer geological model; (c) Case 3a: SimSLE starting with a uniform $K = 8.0 \times 10^{-6}$ m/s; (d) Case 3b: SimSLE using the calibrated 5-layer geological model as prior distribution; (e) Case 3c: SimSLE using the calibrated 19-layer geological model as prior distribution; and (f) Case 3d: SimSLE calibration case using the uncalibrated 19-layer model assigned with permeameter test *K* values for each layer as prior distribution.

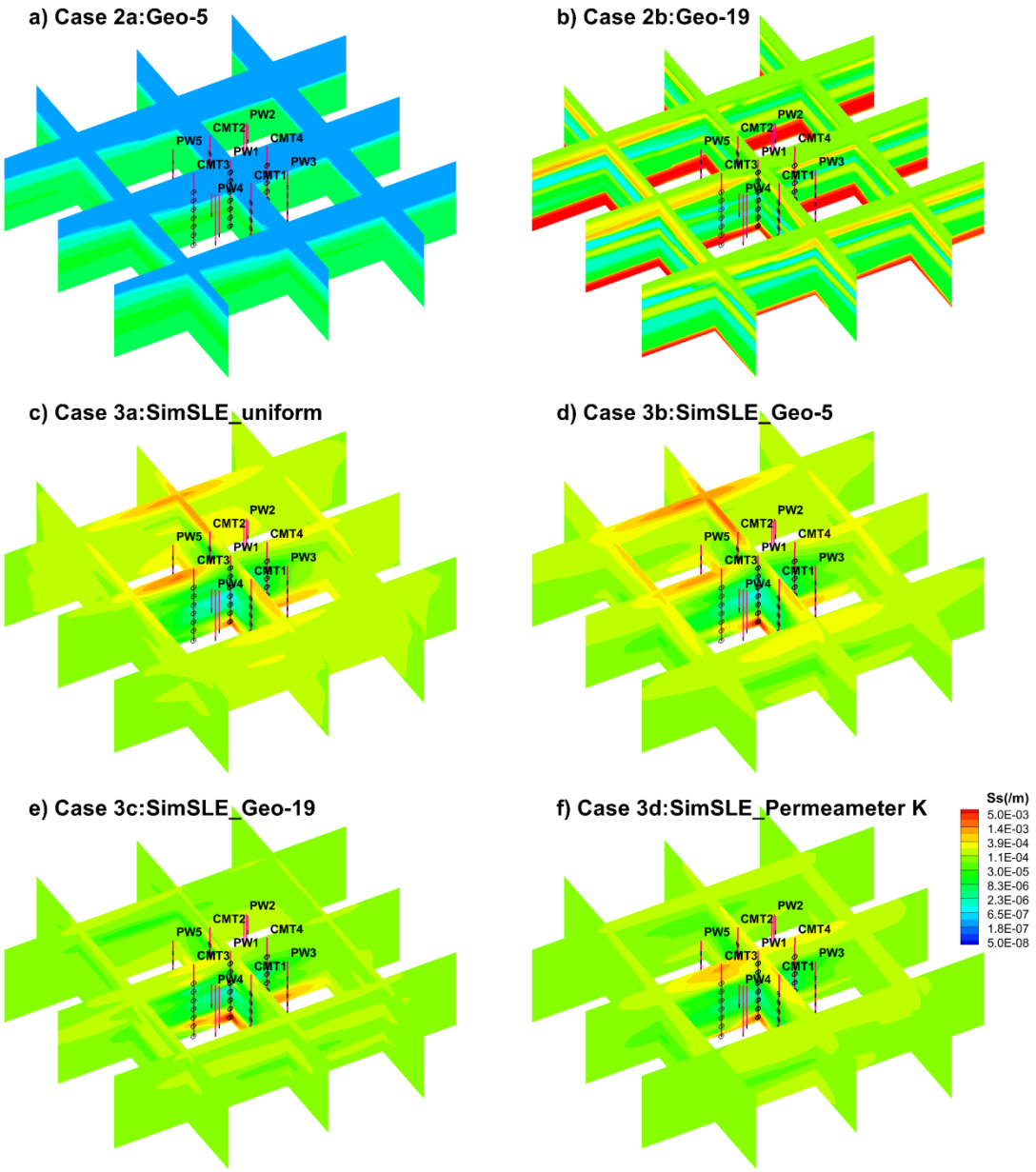


Fig. 4. Estimated S_s fields from the inversion of eight pumping tests for: (a) Case 2a: the 5-layer geological model; (b) Case 2b: the 19-layer geological model; (c) Case 3a: SimSLE starting with a uniform $K = 8.0 \times 10^{-6}$ m/s; (d) Case 3b: SimSLE using the calibrated 5-layer geological model as prior distribution; (e) Case 3c: SimSLE using the calibrated 19-layer geological model as prior distribution; and (f) Case 3d: SimSLE calibration case using the uncalibrated 19-layer model assigned with permeameter test K values for each layer as prior distribution.

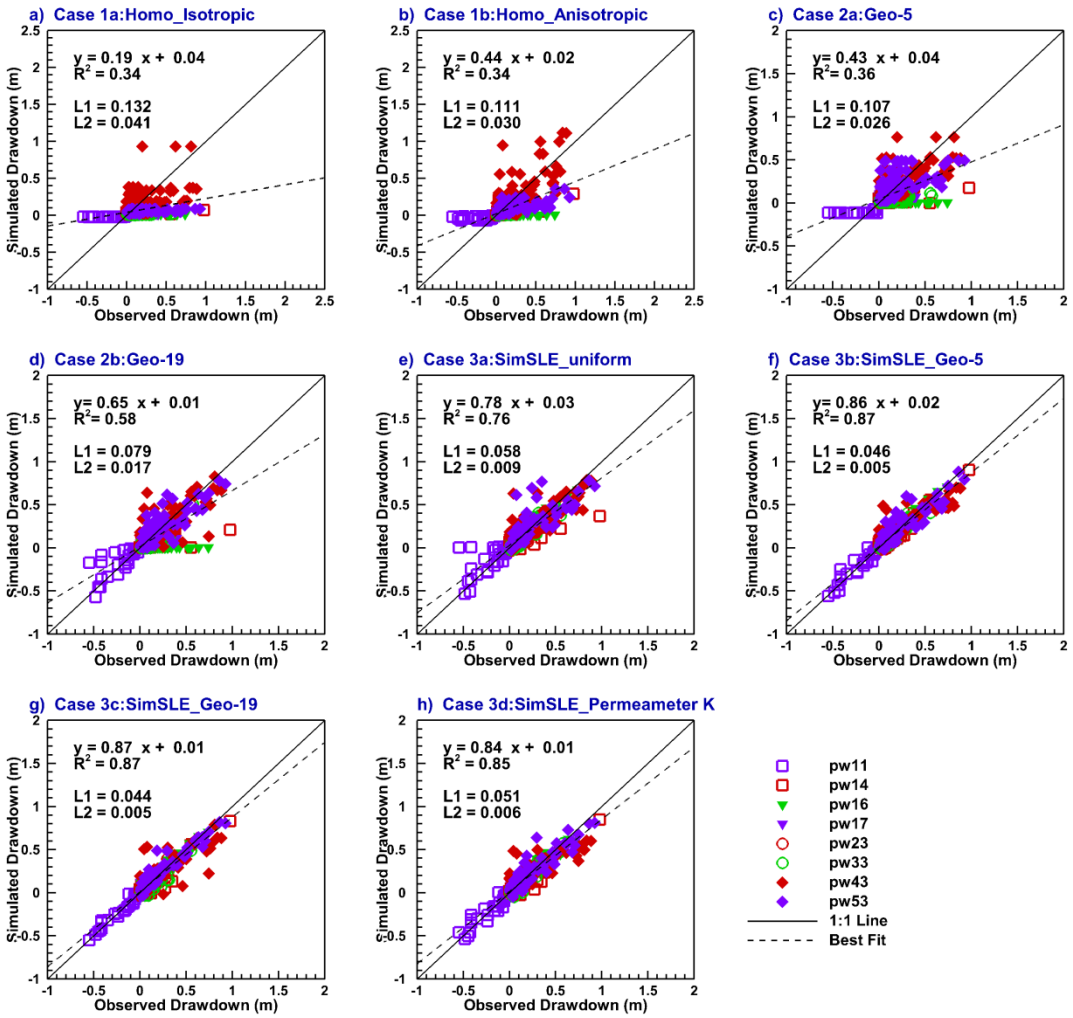


Fig. 5. Scatterplots of observed versus simulated drawdowns for model calibrations using eight pumping tests for the: (a) Case 1a: isotropic effective parameter model; (b) Case 1b: anisotropic effective parameter model; (c) Case 2a: geological model with five layers; (d) Case 2b: geological model with 19 layers; (e) Case 3a: SimSLE starting with $K = 8.0 \times 10^{-6}$ m/s as prior mean; (f) Case 3b: SimSLE using the calibrated five-layer geological model as prior distribution; (g) Case 3c: SimSLE using the calibrated 19-layer geological model as prior distribution; and (h) Case 3d: 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.

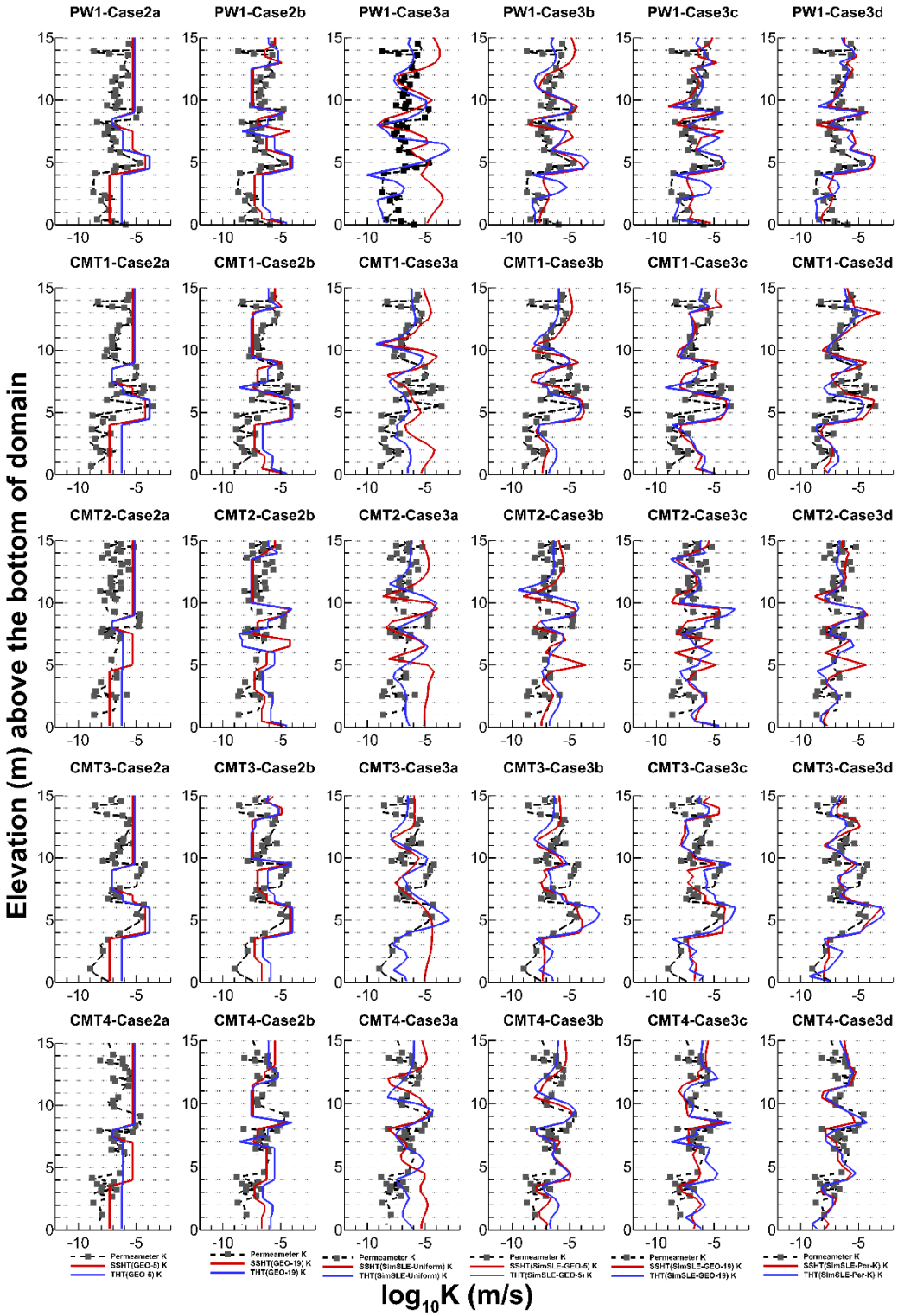
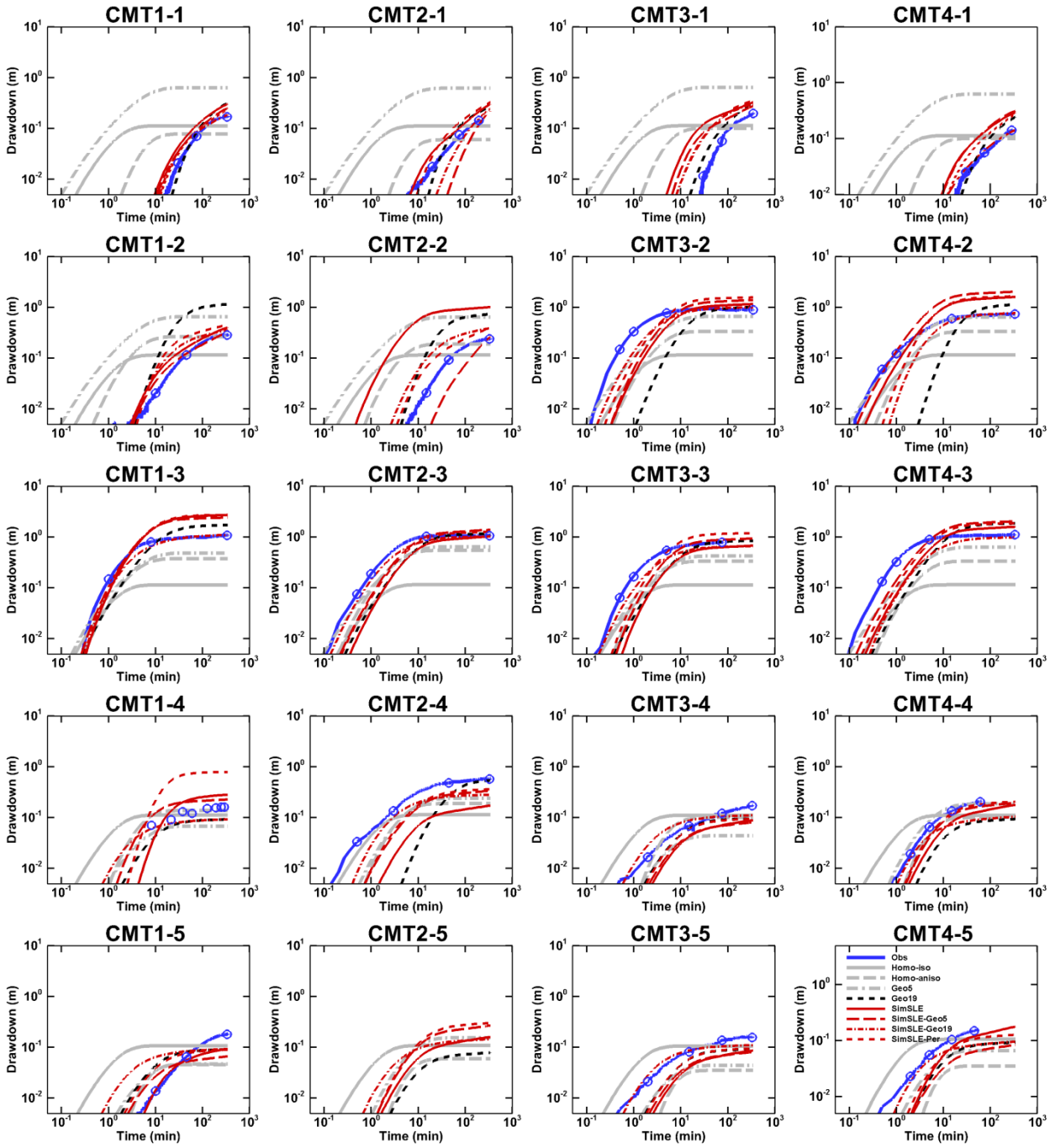


Fig. 6. Vertical $\log_{10}K$ profiles along nine boreholes of PW1 and CMT1-4 wells, for different Cases 2a, 2b, 3a, 3b, 3c and 3d.

(a)



(b)

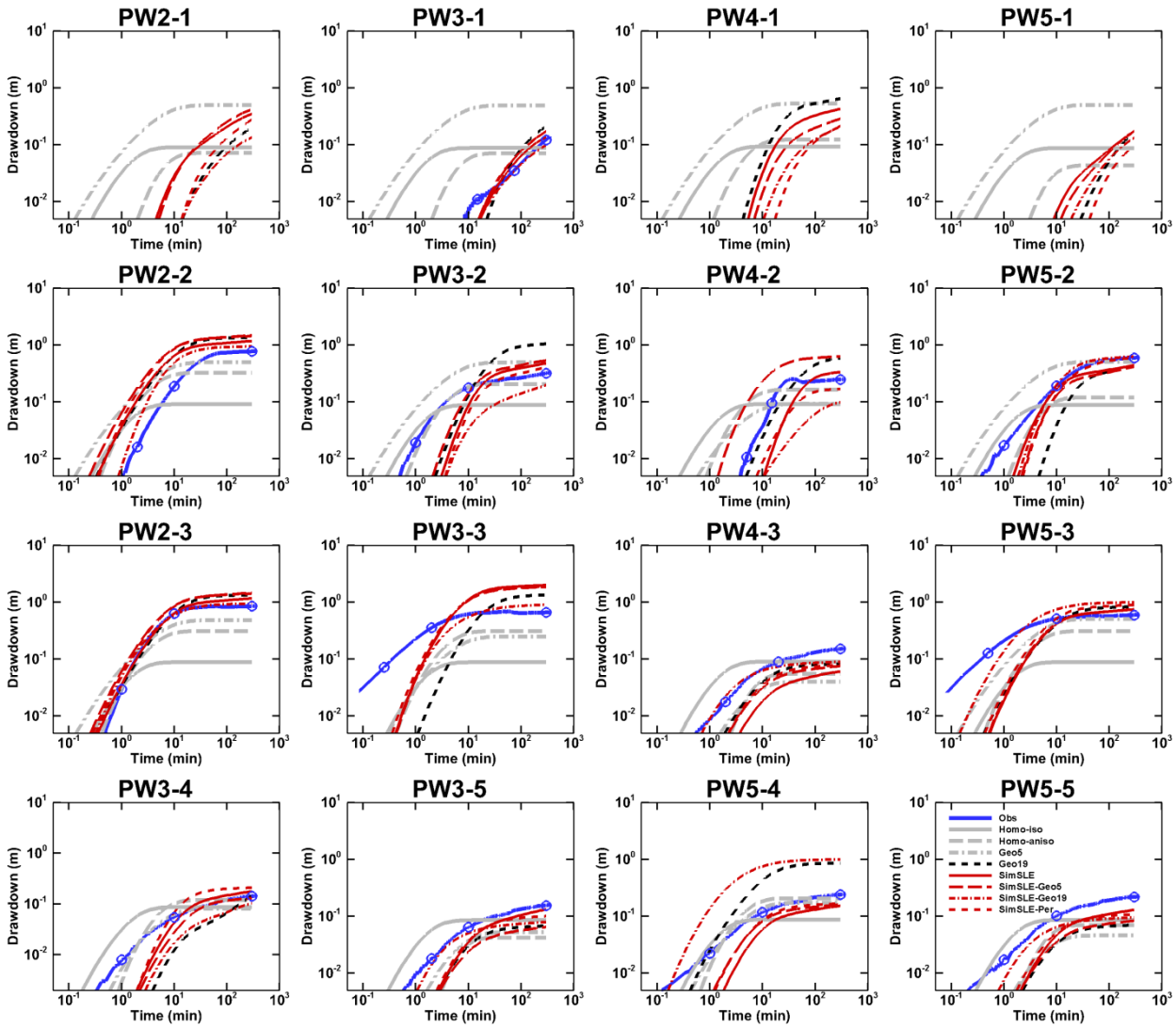


Fig. 7. Observed and simulated drawdowns from validation pumping test at PW1-3 in (a) CMT and (b) PW ports. The solid blue lines represent observed drawdowns; the solid gray lines represent simulated drawdowns by Case 1a; the dashed gray lines represent simulated drawdowns by Case 1b; the dash-dotted gray lines represent simulated drawdowns by Case 2a; the dotted black lines represent simulated drawdowns by Case 2b; the solid red lines represent simulated drawdowns by Case 3a; the dashed red lines represent simulated drawdowns by Case 3b; the dash-dotted red lines represent simulated drawdowns by Case 3c; the dotted red lines represent simulated drawdowns by Case 3d; the blue circles represent time points selected for validation comparisons.

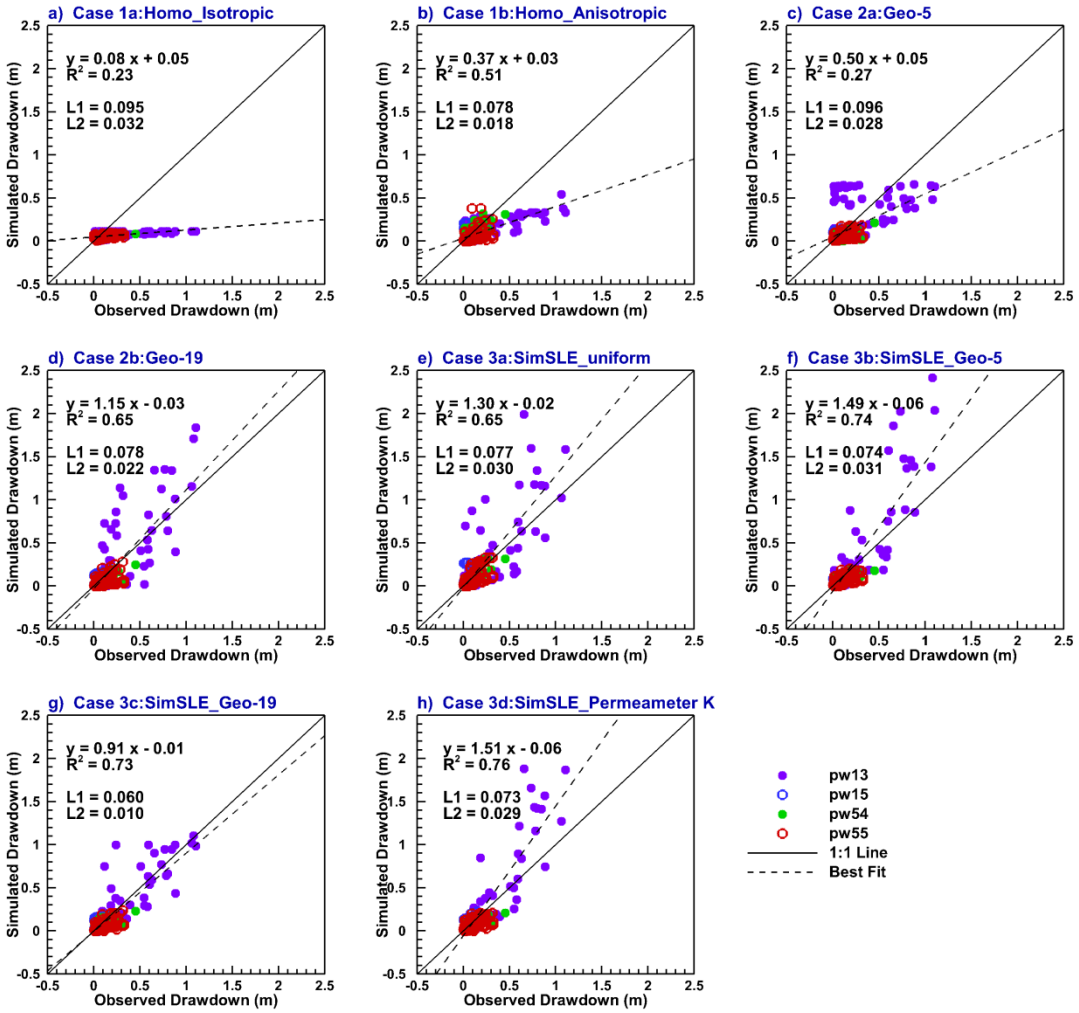


Fig. 8. Scatterplots of observed versus simulated drawdowns for model validation using four pumping tests for the: (a) Case 1a: isotropic effective parameter model; (b) Case 1b: anisotropic effective parameter model; (c) Case 2a: geological model with five layers; (d) Case 2b: geological model with 19 layers; (e) Case 3a: SimSLE starting with $K = 8.0 \times 10^{-6}$ m/s as prior mean; (f) Case 3b: SimSLE using the calibrated five-layer geological model as prior distribution; (g) Case 3c: SimSLE using the calibrated 19-layer geological model as prior distribution; and (h) Case 3d: 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.

Table 1: Estimated K and S_s values as well as their posterior 95% confidence intervals for the effective parameter approach.

	Estimated	95% Confidence Intervals		Estimated	95% Confidence Intervals	
	K (m/s)	Lower limit	Upper limit	S_s (/m)	Lower limit	Upper limit
Case 1a	2.38×10^{-5}	2.06×10^{-5}	2.76×10^{-5}	9.34×10^{-6}	2.56×10^{-6}	3.41×10^{-5}
Case 1b	K_x 1.85×10^{-5}	1.42×10^{-5}	2.42×10^{-5}	1.39×10^{-5}	7.12×10^{-6}	2.72×10^{-5}
	K_y 2.55×10^{-5}	2.04×10^{-5}	3.20×10^{-5}			
	K_z 3.77×10^{-7}	2.73×10^{-7}	5.21×10^{-7}			

Table 2: Estimated K and S_s values as well as their posterior 95% confidence intervals for the 5-layer geological model. Italicized numbers indicate unrealistic confidence interval limits.

Layer	Estimated K (m/s)	95% Confidence Intervals		Estimated S_s (/m)	95% Confidence Intervals	
		Lower limit	Upper limit		Lower limit	Upper limit
1[*a]	7.86×10^{-6}	6.56×10^{-6}	9.42×10^{-6}	2.29×10^{-7}	*7.99×10^{-18}*	*$6.58\times10^{+3}$*
11	8.73×10^{-8}	4.76×10^{-8}	1.60×10^{-7}	1.43×10^{-6}	*4.00×10^{-16}*	*$5.12\times10^{+3}$*
12[*b]	7.21×10^{-7}	2.46×10^{-7}	2.11×10^{-6}	6.09×10^{-6}	1.68×10^{-8}	2.22×10^{-3}
15	1.47×10^{-4}	1.28×10^{-4}	1.69×10^{-4}	9.28×10^{-6}	4.42×10^{-8}	1.95×10^{-3}
16[*c]	5.78×10^{-7}	2.81×10^{-7}	1.19×10^{-6}	7.28×10^{-6}	7.70×10^{-8}	6.89×10^{-4}

[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.

Table 3: Estimated K and S_s values as well as their posterior 95% confidence intervals for the 19-layer geological model. Italicized numbers indicate unrealistic confidence interval limits.

Layer	Estimated K (m/s)	95% Confidence Intervals		Estimated S_s (/m)	95% Confidence Intervals	
		Lower limit	Upper limit		Lower limit	Upper limit
1	9.21×10^{-07}	*7.12×10^{-14}*	*$1.19\times10^{+01}$*	8.14×10^{-05}	*4.60×10^{-12}*	*$1.44\times10^{+03}$*
2	5.77×10^{-06}	*7.18×10^{-12}*	*$4.63\times10^{+00}$*	3.97×10^{-05}	*2.58×10^{-65}*	*$6.12\times10^{+55}$*
3	6.49×10^{-07}	*2.35×10^{-33}*	*$1.79\times10^{+20}$*	2.22×10^{-04}	*2.11×10^{-39}*	*$2.33\times10^{+31}$*
4	6.86×10^{-06}	3.07×10^{-08}	1.53×10^{-03}	3.61×10^{-04}	5.19×10^{-11}	$2.51\times10^{+03}$
5	1.86×10^{-06}	*5.90×10^{-33}*	*$5.85\times10^{+20}$*	1.08×10^{-04}	*1.08×10^{-304}*	*$1.08\times10^{+296}$*
6	6.29×10^{-08}	1.22×10^{-09}	3.26×10^{-06}	7.92×10^{-04}	*4.04×10^{-07}*	*$1.55\times10^{+00}$*
7	2.86×10^{-08}	5.10×10^{-09}	1.61×10^{-07}	2.66×10^{-05}	2.75×10^{-06}	2.57×10^{-04}
8	5.02×10^{-06}	1.42×10^{-06}	1.78×10^{-05}	1.62×10^{-06}	*2.53×10^{-144}*	*$1.04\times10^{+132}$*
9	9.92×10^{-05}	5.85×10^{-05}	1.68×10^{-04}	4.34×10^{-06}	*1.18×10^{-94}*	*$1.59\times10^{+83}$*
10	1.91×10^{-05}	8.89×10^{-06}	4.11×10^{-05}	4.80×10^{-06}	*1.12×10^{-57}*	*$2.05\times10^{+46}$*
11	7.63×10^{-07}	7.63×10^{-07}	7.63×10^{-07}	1.02×10^{-06}	*5.87×10^{-109}*	*$1.77\times10^{+96}$*
12	2.53×10^{-09}	3.08×10^{-10}	2.07×10^{-08}	1.89×10^{-04}	2.38×10^{-05}	1.50×10^{-03}
13	4.71×10^{-09}	6.62×10^{-10}	3.36×10^{-08}	8.59×10^{-06}	9.41×10^{-09}	7.84×10^{-03}
14	3.04×10^{-06}	1.27×10^{-06}	7.30×10^{-06}	1.36×10^{-05}	1.65×10^{-09}	1.12×10^{-01}
15	1.07×10^{-04}	9.13×10^{-05}	1.26×10^{-04}	1.90×10^{-06}	*1.13×10^{-18}*	*$3.19\times10^{+06}$*
16	2.90×10^{-07}	1.25×10^{-07}	6.73×10^{-07}	1.14×10^{-05}	1.07×10^{-09}	1.21×10^{-01}
17	2.10×10^{-06}	2.34×10^{-11}	1.88×10^{-01}	5.01×10^{-03}	*1.11×10^{-08}*	*$2.26\times10^{+03}$*
18	1.42×10^{-06}	4.82×10^{-12}	4.16×10^{-01}	5.88×10^{-03}	1.30×10^{-04}	2.66×10^{-01}
19	3.61×10^{-05}	*1.13×10^{-10}*	*$1.15\times10^{+01}$*	1.22×10^{-04}	*1.22×10^{-304}*	*$1.22\times10^{+296}$*

Highlights

- Long-term transient drawdown data from both aquifer and aquitard layers are used for Transient Hydraulic Tomography (THT) analyses.
- Geostatistical inverse modeling yields the best calibration and validation performances when compared to lower resolution approaches.
- Drawdowns from aquitards improve hydraulic conductivity estimates for the geostatistical inversion approach.
- Reliable geological data are useful for THT analyses at highly heterogeneous sites when pumping/monitoring intervals are sparse.