SAND2016-1277J

Scalable subsurface inverse modeling of huge data sets with an application to tracer concentration breakthrough data from magnetic resonance imaging

Jonghyun Lee[1], Hongkyu Yoon[2], Peter K. Kitanidis[1], Charles J. Werth[3] and Albert J. Valocchi[4]

[1]Department of Civil and Environmental Engineering, Stanford University, Stanford, California, USA.

[2]Geoscience Research and Applications Group, Sandia National Laboratories, Albuquerque, NM 87185, USA

[3]Department of Civil and Environmental Engineering, University of Texas, Austin, Texas, USA.

[4]Department of Civil and Environmental Engineering, University of Illinois, Urbana-Champaign, Illinois, USA.

Abstract. Characterizing subsurface properties is crucial for reliable and cost-effective groundwater supply management and contaminant remediation. With recent advances in sensor technology, large volumes of hydro-geophysical and geochemical data can be obtained to achieve high-resolution images of subsurface properties. However, characterization with such a large amount of information requires prohibitive computational costs associated with "big data" processing and numerous large-scale numerical simulations. To tackle such difficulties, the Principal Component Geostatistical Approach (PCGA) has been proposed as a "Jacobian-free" inversion method that requires κ forward simulation runs for each iteration where κ is much smaller than the number of unknown parameters and measurements. PCGA can be conveniently linked to any multi-physics simulation software with independent parallel executions. In this paper, we extend PCGA to handle a large number of measurements (e.g. 10^6 or more) by constructing a fast preconditioner whose computational cost scales linearly with the data size. For illustration, we characterize the heterogeneous hydraulic conductivity (K) distribution in a laboratory-scale 3-D sand box using about 6 million transient tracer concentration measurements obtained using magnetic resonance imaging. Since each individual observation has little information on the K distribution, the data was compressed by the zero-th temporal moment of breakthrough curves, which is equivalent to the mean travel time under the experimental setting. Only about 2,000 forward simulations in total were required to obtain the best estimate with corresponding estimation uncertainty, and the estimated K field cap-

tured key patterns of the original packing design, showing the efficiency and effectiveness of the proposed method.

1. Introduction

Typical subsurface inverse problems deal with the estimation of geologic heterogeneous parameters, such as hydraulic conductivity, from noisy and sparse measurements including hydraulic head, solute concentration, temperature and so on. It is well known that subsurface inverse problems are underdetermined and ill-posed; that is, the solution to the inverse problem is non-unique and sensitive to measurement and conceptual modeling errors. Therefore, the solution to the inverse problem and its uncertainty are evaluated within a statistical framework. [*McLaughlin and Townley*, 1996; *Carrera et al.*, 2005; *Oliver et al.*, 2008; *Kaipio and Somersalo*, 2007; *Stuart*, 2010; *Smith*, 2014].

With recent advances in sensor and computation technology, unprecedented large volumes of hydro-geophysical and geochemical data sets can be obtained and processed [*Hampson et al.*, 2008; *Barnhart et al.*, 2010; *Orellana and Haigh*, 2008; *Pamukcu and Ghazanfari*, 2014] to achieve high-resolution images of subsurface properties for more accurate and reliable subsurface flow and reactive transport prediction. While a large data set may yield richer and more revealing information to improve inversion results and reduce estimation uncertainty, incorporating a plethora of information into subsurface characterization requires high, often prohibitive, computational costs associated with "big data" processing and a large number of high-dimensional, coupled multi-physics numerical simulations. For example, in a recent extensive hydraulic tomography campaign [*Hochstetler et al.*, 2015], millions of transient pressure data were successfully acquired from a field site, but only a few thousand measurements were carefully selected and used

to perform high-resolution 3-D transient hydraulic tomography to reduce computational costs.

Traditional inversion techniques are not well suited for high-dimensional and joint inverse problems with massive datasets because they usually require a number of numerical simulation model runs proportional to the size of unknowns and measurements in order to construct Jacobian (i.e., *sensitivity*) matrices. To avoid a large number of expensive simulations, a Newton-conjugate gradient (Newton-CG) type method [*Haber and Ascher*, 2001; *Epanomeritakis et al.*, 2008] for nonlinear least square type inversion has been applied using inner conjugate gradient iterations that avoid full Hessian products by forming Hessian-vector products followed by outer Gauss-Newton iterations. While an inner conjugate gradient iteration requires only a few forward and adjoint system solutions independent of the problem size, the entire inversion may require a large number of sequential inner and outer iterations to converge without a good preconditioner.

Another effective approach dealing with massive data is to use a small number of summarized or subsampled data from the original data set in order to save computation and storage costs. When the information content of an individual data record is low, with local influence, and/or redundant with other records, inversions using the entire or reduced data set often provide similar estimation results with comparable uncertainty reduction. In hydrogeology, temporal moments of the large data set such as transient pressure [*Zhu and Yeh*, 2006; *Yin and Illman*, 2009] and concentration breakthrough curves [*Harvey and Gorelick*, 1995a; *Cirpka and Kitanidis*, 2000; *Nowak and Cirpka*, 2006] are widely used to reconstruct unknown hydraulic conductivity fields. Randomized dimensionality reduction methods [*Krebs et al.*, 2009; *Haber et al.*, 2012; *Aravkin et al.*, 2012] using a

random subset of data have been actively studied to achieve an acceptable inversion result close to the best estimate obtained from the complete data set.

Among recently proposed scalable inversion approaches, the Principal Component Geostatistical Approach (PCGA) [*Lee and Kitanidis*, 2014; *Kitanidis and Lee*, 2014] is an approximate method to the Bayesian geostatistical approach [*Kitanidis*, 1995, 2010] to estimate unknown subsurface spatial parameters and quantify the corresponding uncertainty rigorously with an affordable number of forward simulations independent of the problem size. PCGA has been applied to several engineering applications with high dimensional unknown parameters such as hydraulic tomography, tracer data inversion [*Lee and Kitanidis*, 2014], deep aquifer characterization with heat tracer [*Lee et al.*, 2015] and arsenic-bearing mineral imaging [*Fakhreddine et al.*, 2015]. However, PCGA has not been applied to inverse problems with a large number of measurements.

In this paper, we extend PCGA to handle a large number of measurements ($\sim \mathcal{O}(10^6)$), an exercise that will soon become routine in the era of big data. To handle the large cokriging matrix arising from the geostatistical approach, a scalable and exact preconditioner for PCGA is constructed. By scalability, we mean the ability of PCGA to deal with very large measurements and unknowns, and the computation/storage costs of the proposed preconditioner increase linearly with respect to the dimension of measurements and unknowns. Our proposed method is used to estimate the unknown hydraulic conductivity field in a 3-D laboratory-scale sand box from in-situ tracer breakthrough data obtained using magnetic resonance imaging (MRI) [*Yoon et al.*, 2008]. The same data set was used for the sand box characterization by *Yoon and McKenna* [2012]. The previous work implemented PEST [*Doherty and Hunt*, 2010] linked with MODFLOW [*Harbaugh et al.*,

2000], and an advective particle tracking method instead of the full advection-dispersion simulation model due to the prohibitive computational costs. In this work, we use coupled flow and transport simulation with MODFLOW [*Harbaugh et al.*, 2000] and MT3DMS [*Zheng and Wang*, 1999], respectively, to achieve a (approximate) full geostatistical solution, and also corresponding solution uncertainty that was not reported in the previous work.

This paper is organized in the following way. Section 2 reviews the geostatistical and principal component geostatistical approaches. The computational framework for large data set inversion is also presented. Section 3 explains the MRI experiment setup briefly, and data reduction technique. In Section 4, PCGA is applied to two synthetic examples to investigate the computational efficiency and solution accuracy of the proposed method. Then, inversion results using the real experimental data set are presented. Concluding remarks follow in Section 5.

2. Method

In this section, we review the quasi-linear geostatistical approach [*Kitanidis*, 1995] and PCGA [*Kitanidis and Lee*, 2014; *Lee and Kitanidis*, 2014]. Then we extend PCGA to solve large data inversion problems.

2.1. Review of Geostatistical Approach

The observation equation, which relates the $m \times 1$ vector of unknowns $\mathbf{s}$ to the $n \times 1$ vector of the data $\mathbf{y}$ is

$$\mathbf{y} = h(\mathbf{s}) + \mathbf{v}, \quad \mathbf{v} \sim N(0, \mathbf{R}) \tag{1}$$

where h is the forward model mapping the parameter space $\mathbb{R}^m$ to the measurement space $\mathbb{R}^n$, $\mathbf{v}$ is Gaussian with zero mean and covariance $\mathbf{R}$ that accounts for errors in the data $\mathbf{y}$ and the forward model h. The prior probability of $\mathbf{s}$ is Gaussian with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance $\mathbf{Q}$, where $\mathbf{X}$ is the $m \times p$ known (polynomial) matrix, $\boldsymbol{\beta}$ is the $p \times 1$ unknown vector (typically p = 1), and $\mathbf{Q}$ is the generalized covariance matrix [*Kitanidis*, 1983, 1993].

The posterior pdf of $\mathbf{s}$ and $\boldsymbol{\beta}$ are obtained through Bayes theorem and its negative loglikelihood, $-\ln p''(\mathbf{s}, \boldsymbol{\beta})$, is

$$-\ln p''(\mathbf{s}, \boldsymbol{\beta}) = \frac{1}{2}(\mathbf{y} - \mathbf{h}(\mathbf{s}))^\top \mathbf{R}^{-1}(\mathbf{y} - \mathbf{h}(\mathbf{s})) + \frac{1}{2}(\mathbf{s} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{Q}^{-1}(\mathbf{s} - \mathbf{X}\boldsymbol{\beta}) \tag{2}$$

By minimizing (2) with respect to $\mathbf{s}$ and $\boldsymbol{\beta}$, we can obtain the maximum a posterior (MAP) or most likely value $\hat{\mathbf{s}}$, commonly computed through an iterative Gaussian-Newton method.

For this method, we start with the latest "best" estimate $\bar{\mathbf{s}}_i$, and update to a new solution $\bar{\mathbf{s}}_{i+1}$. Next, the $n \times m$ Jacobian or senstivity matrix $\mathbf{H}$ of $\mathbf{h}$ at $\bar{\mathbf{s}}_{i+1}$ is evaluated as:

$$\mathbf{H} = \left.\frac{\partial \mathbf{h}}{\partial \mathbf{s}}\right|_{\mathbf{s}=\bar{\mathbf{s}}_i} \tag{3}$$

Then, based on the linearization of (2), the updated solution for the next iteration is computed as

$$\bar{\mathbf{s}}_{i+1} = \mathbf{X}\bar{\boldsymbol{\beta}} + \mathbf{Q}\mathbf{H}^\top \bar{\boldsymbol{\xi}} \tag{4}$$

where $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\xi}}$ are computed by solving a single linear system of $n + p$ equations:

$$\begin{bmatrix} \mathbf{H}\mathbf{Q}\mathbf{H}^\top + \mathbf{R} & \mathbf{H}\mathbf{X} \\ (\mathbf{H}\mathbf{X})^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{\xi}} \\ \bar{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{y} - h(\bar{\mathbf{s}}_i) + \mathbf{H}\bar{\mathbf{s}}_i \\ \mathbf{0} \end{bmatrix} \tag{5}$$

Steps (3) - (5) are repeated until $\bar{\mathbf{s}}_i$ converges to the best estimate $\hat{\mathbf{s}}$. For strongly nonlinear problems, a Levenberg-Marquardt type method using a larger error matrix $\mathbf{R}_\alpha = \alpha\mathbf{R}$ (where $\alpha \geq 1$) adaptively for the first few iterations can be used for better convergence. Note that the computation of the Jacobian requires $\min(m, n) + 1$ forward simulation runs using forward or adjoint-state method. Once $\hat{\mathbf{s}}$ is obtained, the posterior covariance matrix $\mathbf{V}$ is computed as

$$\mathbf{V} = \mathbf{Q} - \begin{bmatrix} \mathbf{HQ} \\ \mathbf{X}^\top \end{bmatrix}^\top \begin{bmatrix} \mathbf{HQH}^\top + \mathbf{R} & \mathbf{HX} \\ (\mathbf{HX})^\top & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{HQ} \\ \mathbf{X}^\top \end{bmatrix} \tag{6}$$

2.2. Principal Component Geostatistical Approach

While the geostatistical method is well suited for small- to moderate-scale inverse problems, computational cost can become extremely high when the method is implemented on finely resolved grid with a large number of measurements. The challenges originate from the construction of Jacobian $\mathbf{H}$ and the matrix products of Jacobian, particularly $\mathbf{HQ}$. Separate construction of $\mathbf{H}$ and products of $\mathbf{H}$ with dense matrices lead to at least $n+1$ forward simulations at each iteration and $\mathcal{O}(m^2 n)$ multiplication and $\mathcal{O}(mn)$ storage costs. Furthermore, computation of $\mathbf{H}$ typically requires intrusive changes in the forward model code, which adds another level of difficulty, especially for multi-physics problems that utilize multiple forward models, in series or coupled.

PCGA expedites the geostatistical approach by avoiding the direct evaluation of the Jacobian, by using 1) a low-rank approximation of the prior covariance $\mathbf{Q}$ and 2) a finite-difference approximation of matrix products. Assume $\mathbf{Q}$ is approximated through rank-κ truncated eigen-decomposition:

$$\mathbf{Q} \approx \mathbf{Q}_\kappa = \boldsymbol{Z}\boldsymbol{Z}^\top = \sum_{i=1}^{\kappa} \zeta_i \zeta_i^\top \tag{7}$$

where $\mathbf{Q}_\kappa$ is the rank-κ ($\kappa \ll m$) approximation of $\mathbf{Q}$, $\boldsymbol{Z}$ is the square root of $\mathbf{Q}_\kappa$ using the eigen-decompsion, and ζ_i is the i-th column vector of $\boldsymbol{Z}$ which is the i-th (largest) eigenvector multiplied by the square root of the corresponding i-th eigenvalue of $\mathbf{Q}$. A fast and scalable method to obtain (7) for large-scale covariance matrices is explained in *Lee and Kitanidis* [2014].

A generic Jacobian-vector product $\mathbf{Hu}$ needed in (5) (*e.g.,* $\mathbf{u} = \bar{\mathbf{s}}$, $\mathbf{X}_i$ or $\boldsymbol{\zeta}_i$) can be computed approximately at the cost of an additional forward model evaluation using a finite-difference approximation:

$$\mathbf{Hu} = \frac{1}{\delta}\left[h(\mathbf{u} + \delta\mathbf{u}) - h(\mathbf{u})\right] + \mathcal{O}(\delta) \tag{8}$$

where δ is the finite-difference perturbation size. An optimal choice of δ [*Brown and Saad*, 1990] is given by

$$\hat{\delta} = \frac{\sqrt{\epsilon}}{\|\mathbf{u}\|_2^2} \max\left(|\mathbf{s}^\top\mathbf{u}|, \mathbf{c}^\top|\mathbf{u}|\right) \operatorname{sign}\left(\mathbf{s}^\top\mathbf{u}\right) \tag{9}$$

where ϵ is the relative machine precision, which is usually (one order of magnitude) greater than the square root of the machine precision, $|\mathbf{u}| = [|u_1|, |u_2|, ..., |u_m|]^\top$, $\mathbf{c} = [c_1, c_2, ...c_m]^\top$, c_i is a typical value of $|s_i|$ and sign() indicates a sign of value.

Accordingly, the matrix-matrix products $\mathbf{HQ}$ and $\mathbf{HQH}^\top$ are computed by

$$\mathbf{HQ} \approx \mathbf{HQ}_\kappa = \mathbf{H}\sum_{i=1}^{\kappa}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top = \sum_{i=1}^{\kappa}\left(\mathbf{H}\boldsymbol{\zeta}_i\right)\boldsymbol{\zeta}_i^\top \approx \sum_{i=1}^{\kappa}\boldsymbol{\eta}_i\boldsymbol{\zeta}_i^\top \tag{10}$$

$$\mathbf{HQH}^\top \approx \mathbf{HQ}_\kappa\mathbf{H}^\top = \sum_{i=1}^{\kappa}\left(\mathbf{H}\boldsymbol{\zeta}_i\right)\left(\mathbf{H}\boldsymbol{\zeta}_i\right)^\top \approx \sum_{i=1}^{\kappa}\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\top \tag{11}$$

where

$$\boldsymbol{\eta}_i = \mathbf{H}\boldsymbol{\zeta}_i \approx \frac{1}{\delta}\left[h(\mathbf{s} + \delta\boldsymbol{\zeta}_i) - h(\mathbf{s})\right] \tag{12}$$

.

Replacing the explicit construction and multiplication of $\mathbf{H}$ in (3) - (5) by (8) - (12) requires a total of $\kappa + p + 2$ forward model runs in each iteration, and the storage cost becomes $\mathcal{O}(m\kappa)$ from $\mathcal{O}(mn)$. As a result, PCGA can reduce the number of numerical simulations significantly from $n+1$ to $\kappa+p+2$ when a large number of measurements are available. Previous numerical experiments [*Lee and Kitanidis*, 2014; *Fakhreddine et al.*, 2015; *Lee et al.*, 2015] have shown that $\kappa \sim \mathcal{O}(100)$ and a few hundred simulation runs in total are needed without any intrusive changes in the simulation model code, while inverse solutions are almost the same as those obtained from the geostatistical approach. Corresponding estimation variance can be efficiently computed as in Appendix B.

2.3. Fast and Exact Preconditioner for PCGA

Our previous research presented high-dimensional and/or joint inversion problems with a moderate number of measurements ($\sim \mathcal{O}(10^3)$). When a massive data set is available, solving the $n + p$ by $n + p$ cokriging system in (13), for example $n = 10^6$, would be infeasible with direct matrix inversion methods and should be implemented with iterative methods such as MINRES [*Paige and Saunders*, 1975] and GMRES [*Saad and Schultz*, 1986]. Those iterative methods usually require a preconditioner to reduce the number of iterations and constructing a "good" preconditioner, which is close to the inverse of the cokriging matrix and guarantees a few iterations, is typically expensive. However, for PCGA we can accelerate the direct solution of (5) or construct an effective preconditioner based on the exact inverse of the cokriging or so-called saddle point matrix [*Benzi et al.*, 2005]:

$$\begin{bmatrix} \mathbf{HQH}^\top + \mathbf{R} & \mathbf{HX} \\ (\mathbf{HX})^\top & \mathbf{0} \end{bmatrix}^{-1} := \begin{bmatrix} \mathbf{\Psi} & \mathbf{\Phi} \\ \mathbf{\Phi}^\top & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1}\mathbf{\Phi}\mathbf{S}^{-1}\mathbf{\Phi}^\top\mathbf{\Psi}^{-1} & \mathbf{\Psi}^{-1}\mathbf{\Phi}\mathbf{S}^{-1} \\ \mathbf{S}^{-1}\mathbf{\Phi}^\top\mathbf{\Psi}^{-1} & -\mathbf{S} \end{bmatrix} \quad (13)$$

where

$$\boldsymbol{\Psi} := \mathbf{HQH}^\top + \mathbf{R}, \quad \boldsymbol{\Phi} := \mathbf{HX}, \quad \mathbf{S} := \boldsymbol{\Phi}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\Phi} \tag{14}$$

and the dominant cost for (13) is the computation of $\boldsymbol{\Psi}^{-1}$, i.e.,

$$\boldsymbol{\Psi}^{-1} = \left(\mathbf{HQH}^\top + \mathbf{R}\right)^{-1} \tag{15}$$

This type of matrix form has been actively researched as an application of the generalized eigenvalue problem (GEP) [*Flath et al.*, 2011; *Cui et al.*, 2014; *Saibaba and Kitanidis*, 2015] to approximate the Hessian of (2) with a relatively small number of terms (generally $\sim \mathcal{O}(100)$); we follow this technique to solve (5) for PCGA. Assume that we solve the following GEP (see Appendix A) to find $\mathbf{u}$ and λ, which are the generalized eigenvector and eigenvalue of $\mathbf{Q}$ and $\mathbf{R}$, respectively:

$$\mathbf{HQH}^\top \mathbf{u} = \lambda \mathbf{R} \mathbf{u} \tag{16}$$

that satisfies

$$\mathbf{HQH}^\top = \mathbf{RU} \boldsymbol{\Lambda} \mathbf{U}^\top \mathbf{R}, \quad \mathbf{URU}^\top = \mathbf{I} \tag{17}$$

where the columns of $\mathbf{U}$ and diagonal values of $\boldsymbol{\Lambda}$ are generalized eigenvectors $\mathbf{u}$ and eigenvalues λ. Then, using the Sherman-Morrison-Woodbury formula,

$$\boldsymbol{\Psi}^{-1} = \left(\mathbf{HQH}^\top + \mathbf{R}\right)^{-1} = \mathbf{R}^{-1} - \mathbf{UDU}^\top \tag{18}$$

where diagonal matrix $\mathbf{D}$ whose i-th diagonal value $\mathbf{D}_i$ is

$$\mathbf{D}_i = \frac{\lambda_i}{\lambda_i + 1} \tag{19}$$

In the PCGA framework, where $\mathbf{Q}$ is replaced with the rank-κ approximation $\mathbf{Q}_\kappa$, we need to find only "κ" generalized eigenmodes for (16) to obtain the exact inverse of the

cokriging matrix used in PCGA using (18):

$$\boldsymbol{\Psi}^{-1} \approx \boldsymbol{\Psi}_{\kappa}^{-1} = \left(\mathbf{H}\mathbf{Q}_{\kappa}\mathbf{H}^{\top} + \mathbf{R}\right)^{-1} = \mathbf{R}^{-1} - \mathbf{U}_{\kappa}\mathbf{D}_{\kappa}\mathbf{U}_{\kappa}^{\top} \tag{20}$$

where $\mathbf{U}_{\kappa}$ and $\boldsymbol{\Lambda}_{\kappa}$ are generalized eigenvectors and eigenvalues of $\mathbf{Q}_{\kappa}$ and $\mathbf{R}$, respectively. GEP can be solved efficiently using the sequential Lanczos-based method or parallelized randomized SVD [*Saibaba et al.*, 2015], with computational cost of $\mathcal{O}(n\kappa^2)$ and storage cost of $\mathcal{O}(n\kappa)$, i.e., linear scalability with respect to the data size. Once the generalized eigenvalues and eigenvectors are computed, the inverse matrix (13) can be used to solve the cokriging system directly or as a preconditioner for iterative approaches.

It should be noted that the iterative approaches require a form of matrix-vector product instead of explicit construction of (13). For example, $\boldsymbol{\Psi}_{\kappa}^{-1}\mathbf{x}$ can be computed as

$$\boldsymbol{\Psi}^{-1}\mathbf{x} \approx \boldsymbol{\Psi}_{\kappa}^{-1}\mathbf{x} = \mathbf{R}^{-1}\mathbf{x} - \mathbf{U}_{\kappa}(\mathbf{D}_{\kappa}(\mathbf{U}_{\kappa}^{\top}\mathbf{x})) \tag{21}$$

without storing and computing the full matrix $\boldsymbol{\Psi}_{\kappa}^{-1}$. The same argument is applied to $\mathbf{S}$ and $\boldsymbol{\Phi}$ in (13). In the numerical experiment we present in the next section with $n = 51,584$ and $\kappa = 2000$, GEP was solved in 10 seconds and MINRES or GMRES required no more than 4 iterations to achieve convergence with a small residual, e.g., 10^{-8}.

2.4. Choice of the number of Principal Components κ

In our previous works [*Kitanidis and Lee*, 2014; *Lee and Kitanidis*, 2014], two methods were proposed to choose a reasonable κ based on 1) relative eigenvalue error (the ratio of $\kappa + 1$-th eigenvalue to the first eigenvalue) of the prior covariance approximation and 2) eigenspectrum of $\mathbf{H}\mathbf{Q}_{\kappa}\mathbf{H}^{\top}$ compared to $\mathbf{R}$. In many practical cases, the former would work effectively by keeping κ principal components that give a small relative eigenvalue error (e.g. ≤ 0.01). However, this criteria alone might not be sufficient, especially when

the information content from dense measurements is rich enough to recover small-scale features, and the eigenvalue decay of the prior $\mathbf{Q}$ chosen for such a data-intensive case is slow due to short correlation length, high physical domain dimension (i.e., 3-D) and high parameter dimension [*Frauenfelder et al.*, 2005]. As a result, κ can vary dramatically depending on the relative eigenvalue error one allows. A more rigorous way would be to investigate the combined effect of the prior information, forward model prediction, and modeling and measurement errors on the approximation of the estimation results. In fact, the generalized eigenvalue of $\mathbf{HQH}^\top$ and $\mathbf{R}$ discussed in Section 2.3 can be an effective mathematical tool for this purpose and interpreted as a generalized Rayleigh-Ritz ratio:

$$\lambda_i = \underset{\mathbf{u}\in\mathbf{U}_{i,i+1,\cdots,n}}{\arg\max} \frac{\mathbf{u}^\top\mathbf{HQH}^\top\mathbf{u}}{\mathbf{u}^\top\mathbf{Ru}} \tag{22}$$

where $\mathbf{U}_{i,i+1,\cdots,n}$ is the space spanned by the eigenvectors corresponding to the eigenvalues equal to or smaller than the i-th eigenvalue. The i-th eigenvalue maximizes the Rayleigh-Ritz ratio over the measurement subspace $\mathbf{U}_{i,i+1,\cdots,n}$, and can be interpreted as how much the unknown $\mathbf{s}$ contributes to the measurement variability compared to the noise $\mathbf{R}$ in the corresponding eigenspace. In other words, if λ_i is greater than 1, the measurements are more important and informative to the solution than the noise along the corresponding eigenvector direction (cf. *Cui et al.* [2014] for the Hessian approximation). Thus, we can choose κ whose eigenvalue λ_κ is close to 1. Note that one can allow κ slightly larger than 1 without losing accuracy since the prior $\mathbf{Q}$ includes redundant information on the prior mean structure $\mathbf{X}$. The generalized prior covariance [*Kitanidis and Lee*, 2014] that excludes the effect of the prior mean can be used for more rigorous choice of κ.

It should be noted that the analysis above is based on $\mathbf{HQH}^\top$, and what we compute with κ principal components is κ-rank $\mathbf{HQ}_\kappa\mathbf{H}^\top$. The κ-th generalized eigenvalue λ_κ of

$\mathbf{HQ}_\kappa\mathbf{H}^\top$ (and $\mathbf{R}$) would be much smaller than actual λ_κ of $\mathbf{HQH}^\top$. Thus, one should check whether the eigenspectrum above $\lambda \approx 1$ changes by adding more principal components in order to make sure the eigenspectrum of $\mathbf{HQ}_\kappa\mathbf{H}^\top$ is close to that of $\mathbf{HQH}^\top$ around λ_κ. A practical and efficient implementation is that one starts PCGA with a small value of κ, then increases κ steadily in each iteration up to the value where the generalized eigenspectrum does not vary much (for example the spectrum above $\lambda = 10$).

3. Application to Tracer Data Inversion

We applied the method proposed in Section 2 to a laboratory-scale hydraulic conductivity estimation problem. The experimental procedure and data acquisition are given in detail by *Yoon et al.* [2008] and *Zhang et al.* [2007]. Here, we briefly explain the experimental setup, the data processing with temporal moment analysis and the numerical setting.

3.1. Experimental Setup

The experimental setup of the flowcell is shown in Figure 1. The entire flowcell has dimensions of 21.5 × 9 × 8.5 cm, and is packed with 1 cm cubes of five different sand types described in Table 1. The hydraulic conductivity, porosity and dispersion coefficients of all five sand types were measured independently before packing. The "true" hydraulic conductivity field with five different sand types was generated using the sequential indicator simulation algorithm [*Deutsch and Journel*, 1998] to construct a heterogeneous K field for the central portion of the flowcell (14 × 8 × 8 cm) as in Figure 1 (b). The rest of the flowcell (a 4.5 cm zone adjacent to the inlet, a 3 cm zone adjacent to the outlet, a 0.5 cm thick layer on the bottom and a 0.5 cm thick vertical layer adjacent to

the side walls) was filled with 50/70 sand (lowest hydraulic conductivity). For the tracer test, a constant head was maintained at the inlet and outlet reservoirs, and a nonreactive para-magnetic tracer solution was continuously injected into the initial tracer-free water saturated flowcell until complete breakthrough was observed in the outflow. The total time for the tracer solution to flow through the entire flowcell was about 4 hours.

The signal intensity for the tracer concentration using MRI was obtained and processed at a resolution of $0.25 \times 0.25 \times 0.25$ cm $= 0.016$ cm^3, at a regular sampling interval of 2.17 min over the MR imaging region, which is slightly smaller than the entire heterogeneous region ($14 \times 8 \times 8$). The MRI signal was converted into normalized tracer breakthrough curves (BTCs), i.e., tracer concentration varies between 0 and 1. Measurements from the top 0.25 cm of the heterogeneous region were not used in this study due to decreased imaging accuracy; the actual observation data for the inversion covers from x = 4.5 and 17.5 cm, y = 0.5 to 8.5 cm, and z = 0.5 to 8.25 cm in the central heterogeneous region, and the total number of observed tracer concentration data records is 5,777,408.

3.2. Data Processing with Temporal Moments

For tracer test data inversion, the first normalized temporal moment of tracer concentration BTCs has been widely used [*Cirpka and Kitanidis*, 2000; *Nowak and Cirpka*, 2006]. The first normalized temporal moment of a local BTC due to a pulse-like tracer injection, $m_{1,n}$ [T], is defined as

$$m_{1,n}(x) = \frac{m_1(x)}{m_0(x)} = \frac{\int_0^\infty tC(x,t)dt}{\int_0^\infty C(x,t)dt} \tag{23}$$

where m_i is the i-th temporal moment [T^{i+1}] and C is the dimensionless normalized concentration [-]. $m_{1,n}$ represents the mean travel (arrival) time of the tracer at a monitoring

location. The use of the first temporal moment in the inversion has been shown to be robust and beneficial because the first temporal moment is continuous with respect to the objective function (2), and the transient transport equation (with time-invariant coefficients) can be transformed into a steady-state equation for the pulse-type tracer injection [*Harvey and Gorelick*, 1995b; *Cirpka and Kitanidis*, 2000].

Previous laboratory-scale studies [*Nowak and Cirpka*, 2006; *Yoon and McKenna*, 2012] performed continuous tracer injection tests and computed the first normalized moment of the derivative of BTCs ($C^{'}$), $m^{'}_{1,n}$ [T], in order to approximate the mean travel time for a pulse-like injection:

$$m^{'}_{1,n}(x) = \frac{m'_1(x)}{m'_0(x)} = \frac{\int_0^T tC'(x,t)dt}{\int_0^T C'(x,t)} \tag{24}$$

where T is the duration of the tracer experiment monitoring. However, as noted in *Yin and Illman* [2009] and *Jose et al.* [2004], care must be taken with additional data processing steps such as derivative computation and denoising; otherwise, a significant source of errors can be introduced, resulting in a very low signal to noise ratio and some information content lost. It is also worth noting that the scheme above is more appropriate for the flux measures; MRI measures the residence concentration from a signal over time within the voxel, which is an average concentration over the MRI scanning time within a voxel.

For the tracer test application considered in this study, we use the relationship between moments of C and moments of $C^{'}$ [*Valocchi*, 1986], and instead of (24), evaluate the zero-th temporal moments, m_0 [T], of the data:

$$m_0(x) = \int_0^T C(x,t)dt \tag{25}$$

The zero-th temporal moment at a particular location represents the total tracer mass that passes by the location of observation and is computed as the area under the BTC. The zero-th temporal moment is equivalent to $m'_{1,n}(x)$ in (24) when the tracer concentration reaches steady state (i.e., $C(x,T)=1$):

$$m'_{1,n}(x) = \frac{m'_1(x)}{m'_0(x)} = \frac{\int_0^T tC'(x,t)dt}{\int_0^T C'(x,t)} = \frac{tC(x,t)|_0^T - \int_0^T C(x,t)dt}{C(x,t)|_0^T} = T - m_0(x) \qquad (26)$$

Thus, in the case of continuous injection, the time, T, to reach steady-state minus the zero-th moment m_0 is the mean travel time of the tracer. Using the zero-th moment is preferable since one can avoid the derivative computation with additional numerical errors. In this work, the zero-th moments were computed from the MRI dataset using the trapezoidal rule to obtain 51,584 mean travel time measurements.

3.3. Numerical Setting

Coupled steady-state flow and transient transport for the tracer test were simulated using USGS MODFLOW [*Harbaugh et al.*, 2000] and MT3DMS [*Zheng and Wang*, 1999]. A uniform grid spacing of 0.25 cm was chosen in the 3-D domain as used in previous works [*Yoon et al.*, 2008; *Yoon and McKenna*, 2012] to have the same scale as the MRI data. A third-order total-variation-diminishing (TVD) scheme was chosen in MT3DMS to prevent numerical dispersion and oscillation. Based on a previous analysis that K is the most sensitive parameter and the travel-time data are relatively insensitive to the dispersivities [*Yoon and McKenna*, 2012; *Nowak and Cirpka*, 2006], only K is estimated in this study while porosity and dispersivity are assumed to be known from measurements taken before packing. Simulation parameters including the domain size, porosity and dispersivity are presented in Table 2.

For parameter estimation, we first generate two synthetic tracer test data sets, each corresponding to a different hypothetical K distribution for the same laboratory setup described in Section 3.1. The purpose of these synthetic tests is to investigate the performance of the proposed method under ideal conditions using known true K fields and conceptual modeling/experimental errors. Then, we employ our method using the actual tracer travel time data to estimate the unknown K distribution of the flowcell. Inversion parameters used in this study are listed in Table 2. Numerical simulations and inversions were carried out on a Linux workstation equipped with 36 Intel core 3.1 GHz processors and 128 GB RAM.

4. Results

4.1. Application to Synthetic Cases

In this section, we consider inversion tests with two synthetic true K fields to investigate the scalability and effectiveness of our proposed method. Eight horizontal layers of the two K fields are plotted in the first row of Figures 2 (a) and (b), respectively. In Case 1, the true K field for the heterogeneous center region of the flowcell was generated from the log-normal distribution with an exponential covariance kernel. The remaining homogeneous part is assumed to be known exactly in order to simplify the structural (prior) parameter selection such as the prior variance, correlation length and measurement error. In Case 2, the actual design K field packed in the flowcell was chosen to check how much the true field can be reconstructed in this synthetic setting prior to the inversion with the actual data. The same laboratory experimental setup described previously was used to generate 5,777,408 transient concentrations from these two K fields, and 10% noise (i.e., standard deviation of error in the concentration measurements = 0.1 $\times$ maximum concentration)

was added. The zero-th temporal moments of corresponding BTCs were computed to obtain $n = 51{,}584$ mean travel time values. The corresponding simulation parameters are listed in Table 2.

The structural parameters for the prior covariance $\mathbf{Q}$ and the error $\mathbf{R}$ were chosen from the parameters used for the true K field generation for Case 1, and using the cR/Q2 criteria [*Kitanidis*, 1991], which is an optimal structure (hyper) parameter selection method in the Bayesian framework, for Case 2. Optimal structural parameter selection within the PCGA framework is beyond the scope of this paper. For both cases, the initial guess was set to a homogeneous field of the natural logarithm of K (ln K) = 2.5, and the best estimate converged in 4 to 5 iterations depending on inversion parameters. We also performed additional tests with different initial guesses such as ln K = 0 and 5, and all the tests converged to the estimates presented below.

The spectrum of the prior covariance $\mathbf{Q}$ is plotted in Figure 3, showing that the decay of the eigenspectrum is slow due to the use of an exponential covariance kernel with a short correlation length defined in the 3-D space. Thus, it is expected that a large value of κ should be retained to reduce the approximation error of Jacobian-covariance products in (10) and (11). Moreover, to resolve the small-scale variability in the true field, PCGA requires large κ to express the high-frequency components in the estimate. We chose $\kappa = 300$ for Case 1 and 500 for Case 2, with relative eigenvalue errors of the prior covariance approximation (the ratio of the $\kappa + 1$-th eigenvalue to the first/largest eigenvalue) of 0.001 and 0.01, respectively. It is worth noting that the full geostatistical approach would require about $n = 51{,}584$ numerical model evaluations in each iteration. A systematic analysis for the optimal κ selection will be presented later.

The best estimate and corresponding estimation uncertainty of the ln K distribution for Case 1 are shown in the second and third rows of Figure 2 (a), respectively. The best estimate identifies high and low K regions, and reproduces the connectivity patterns observed in the true field. Even though a large measurement error (10% of maximum concentration value) was assumed in this application, the information from approximately 6 million concentration values compensates the large error and yields an accurate estimate of the true field. Since the measurements were collected over almost the entire area ($13 \times 8 \times 8$ cm) for the estimation of the heterogeneous region ($14 \times 8 \times 8$ cm), the posterior estimation variance is reduced uniformly. The fitting between simulated and measured mean travel times is plotted in Figure 4 (a). A total of 1,232 MODFLOW and MT3DMS model runs were required to find the best estimate in around 4 hours using independent 36 core parallel executions. The number of model runs includes $\kappa + p + 2$ Jacobian-vector computations in each iteration, and the evaluation of intermediate solutions identified by the Levenberg-Marquardt method between iterations.

For Case 2, the best estimate and its estimation uncertainty of the ln K distribution are presented in the second and third rows of Figure 2 (b), respectively. While the blocky interfaces of different sand types are blurred due to the large measurement error and Gaussian prior assignment, the best estimate identifies interconnected high K channels as well as other small-scale features successfully. Because of the large number of data points, the Gaussian prior becomes unimportant and the data solely guides the delineation of non-Gaussian patterns. The estimation variance around the MRI scanning volume is low, while a high estimation variance is observed upstream of the heterogeneous region and

near the outlet. The measurement fitting is plotted in Figure 4 (b) and a total of 2,124 forward model runs was required, taking about 7.5 hours.

To investigate the effect of the number of principal components, different values of κ principal components are used to estimate ln K fields in Figures 5 and 6, and compute the estimation variances in Figures 7 and 8. As a reference, we use the estimate with κ = 2,000, which is equivalent to the results that we would have gotten if we used the full geostatistical approach, since increasing the κ above 1,000 did not change the best estimate and its variance. In Case 1, where the true smooth log-normal field is estimated, the best estimate even with $\kappa = 300$ is practically similar to the reference estimate. In Case 2, on the other hand, the best estimate requires more principal components to identify small-scale features, and $\kappa = 500$ is enough to obtain the best estimate similar to the reference solution. Similar to the best estimates, the estimation variance with κ = 300 for Case 1 and $\kappa = 500$ for Case 2 are close to the reference variance map. As noted previously, this high accuracy is obtained with PCGA with only 1,232 and 2,140 forward runs respectively, while the traditional adjoint method-based approach would have required at least 51,585 simulations for each iteration, highlighting the scalability of PCGA.

Figure 9 shows the plot of generalized eigenvalues of $\mathbf{HQH}^\top$ and $\mathbf{R}$ for both cases following the method presented in Section 2.4. The generalized eigenvalues for both cases are evaluated at the best estimate with $\kappa = 2000$. It is shown that $\kappa = 300$ for Case 1 and $\kappa = 500$ for Case 2 are indeed practical choices for these synthetic problems.

In Figure 10, we investigate the effect of κ on the accuracy of the best estimates and the estimation variance in terms of element-wise root-mean-square-error (RMSE):

$$\mathrm{RMSE}(lnK) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(lnK_i^{ref} - lnK_i^{est}\right)^2}, \quad \mathrm{RMSE}(v_i) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\sigma_{ref,i}^2 - \sigma_{est,i}^2\right)^2} \tag{27}$$

where $\ln K_i^{ref}$ and $\sigma_{ref,i}^2$ are the reference estimate and estimation variance in the grid cell i, and $\ln K_i^{est}$ and $\sigma_{est,i}^2$ are the ln K estimate and its estimation variance in the same grid cell, respectively. It is also observed that $\kappa \geq 500$ for Case 1 and $\kappa \geq 600$ for Case 2 give negligible errors to the reference estimate and variance.

4.2. Application to MRI Experimental Data

In this section, the actual mean travel time data set is used to estimate the ln K field of the 3-D flowcell and quantify estimation uncertainty. In the previous studies [*Yoon et al.*, 2008; *Yoon and McKenna*, 2012], it was shown that the advection-dispersion simulation based on the actual design packing could not reproduce the measurements obtained from the actual experiment accurately, potentially due to loose sand packing during the construction, sand mixing at interfaces during the experiments, and various sources of errors from MRI data acquisition, MRI signal post-processing and conceptual modeling setup. Thus, the experimental design packing pattern does not exactly match the "true" packing pattern, and is only used for comparison purposes. We refer readers to *Yoon et al.* [2008] and *Yoon and McKenna* [2012] for a rigorous analysis of parameterization and model selections, and a detailed explanation of the simulation model discrepancy.

To save the computational cost further, we started PCGA with $\kappa = 100$ and increased κ by 150 at each iteration following Section 2.4. By doing so, a total of 1,952 MODFLOW and MT3DMS simulations were required to achieve convergence in 5 iterations. From the

cR/Q2 criteria, the inversion parameters were chosen as described in Case 2 of Table 2, except the use of a slightly larger standard deviation of measurement error ($\sigma_t = 4$) to account for the uncertainty arising from measurement and modeling errors.

The best estimate of the ln K field using the actual travel time data is shown in Figure 11. Overall, the best estimate identifies high and low K zones and their connectivity observed in the design packing pattern, while the small-scale features in the best estimate are not exactly the same as those from the previous synthetic test in Figure 2, possibly due to changes and errors in the experiment and modeling setup as mentioned earlier. Corresponding estimation variance in the third row of Figure 11 indicates the high uncertainty outside the MRI scanning volume, especially near the outlet as expected where the downstream K values cannot be inferred from the upstream tracer information. In Figure 12, the data fitting between the simulated and measured mean travel time is displayed. While measurements are reproduced relatively well, measurement fitting allows more errors compared to the previous synthetic case in Figure 4 (b) based on the inversion parameters we found. The impact of different κ values on the estimation is investigated in Figures 13 and 14, indicating $\kappa = 500$ is enough to achieve a full geostatistical solution for this application.

In Table 3, the quality of estimated K distribution compared to the original packing pattern is assessed by a mapping accuracy evaluation used in *Yoon and McKenna* [2012]. In terms of the mapping accuracy and visual comparison to the design packing pattern, the estimated K in this study is significantly better than the previous result (the third column of Table 3) reported in *Yoon and McKenna* [2012] and even comparable to another previous result (the fourth column of Table 3) that directly used the design packing

boundaries of different sand types as prior information. The relatively low mapping accuracy of high K sands (40/50 and 50/70 sands) with respect to the design packing pattern might be explained by mixing at the interfaces between high and low K sands during the sand packing and the flow experiments, resulting in the extension of low K zones. The superior mapping performance of PCGA may be because PCGA approximates the full geostatistical inverse solution in a better way than the pilot-point method used in PEST. The pilot point method often requires a careful placement of pilot points with respect to measurement locations and the choice of an interpolation scheme from the pilot points to the rest of the model domain also affects the estimation results [*Doherty and Hunt*, 2010]. As a result, the pilot point method often yields estimates different from the geostatistical approach [*Oliver et al.*, 2008]. In addition, while the previous study used a particle tracking based advective transport simulation, we included fully coupled MODFLOW-MT3DMS simulation models in the inversion, which would improve the identification of the original packing pattern further.

The computational cost required in PCGA is also smaller than that in *Yoon and McKenna* [2012] who assigned 1,056 pilot points using the explicit knowledge of the heterogeneous and homogeneous regions. Although inversion-accelerating options supported by PEST including the truncated singular value decomposition (TSVD) and the SVD-assist approach [*Doherty and Hunt*, 2010] were tested with various sets of measured data, all of those approaches required construction of the Jacobian matrix, and the overall number of iterations for convergence was higher (e.g., 10-30 iterations) compared to PCGA. This comparison highlights the advantage of PCGA for large-scale and data intensive inverse problems while maintaining the accuracy close to the full geostatistical approach.

Note that the setup cost for the preconditioner is scalable, meaning that the proposed method can handle millions of measurements without difficulty. To demonstrate the scalability of the preconditioner, we performed an additional inversion with 5,777,408 individual tracer data records. The obtained best estimate and estimation variance using the entire data set are not reported here because they are almost the same as those in Figure 11, i.e., the inversion case using only the travel time data, which indicates higher moments data, would not be informative to improve the results or reduce the uncertainty in this case. With $\kappa = 1000$, the generalized eigen-problem for preconditioner construction was solved using a randomized eigen-solver [*Saibaba et al.*, 2015] in 3 minutes, and MINRES required only 3 iterations to achieve convergence in 30 seconds for each Gauss-Newton iteration. Even with linear scalability with respect to the number of measurements, the computation and storage costs for the preconditioner construction and cokriging matrix inversion might become intractable on a personal computer due to the huge data set, but one can use a smaller κ value such as 250 for the preconditioner construction and obtain the same geostatistical solution with slightly more MINRES or GMRES iterations, e.g., 5 to 10 iterations.

5. Concluding Remarks

In this work, we have improved and adapted PCGA, a scalable inversion method, to compute the best estimate and estimation uncertainty using a huge amount of environmental data. A fast and exact preconditioner for PCGA was presented, and a method of choosing the number of principal components, i.e. κ, based on a generalized eigenvalue analysis was provided. The generalized eigenvalue analysis can be a valuable tool for investigating how the combined information from the prior, data, and forward models,

and associated errors affect the performance of PCGA. Overall, for a high-dimensional inverse problem with a large number of unknowns and data, e.g., m and $n \geq 10^6$, our proposed method requires only about κ ($\ll m, n$)) forward simulation runs for each iteration in order to construct Jacobian products, and the matrix computation and storage costs grow linearly with the number of measurements, n. The entire process in PCGA is thus scalable with respect to the unknown parameter and measurement data dimensions, and can be accelerated further by independent parallel forward model executions.

The efficiency and accuracy of PCGA were demonstrated on a massive MRI data set inversion using coupled flow and transport MODFLOW-MT3DMS models. Since PCGA treats available forward models as a "black box", the linkage of MODFLOW-MT3DMS to PCGA was straightforward. Around 6 millions of the concentration measurements converted from MRI signals were reduced to 57,344 mean travel time data records by a temporal moment computation, and PCGA was applied to invert the travel time data with affordable forward runs, only 1,952 MODFLOW-MT3DMS executions much smaller than those required in traditional inversion methods. Due to high information content from the large data set, the estimated K fields captured key patterns of the original sand packing design with a low estimation uncertainty. The efficient inversion of the real environmental data set presented in this paper shows that PCGA is a promising option for large-scale joint inverse problems and can provide an accurate and scalable estimation of unknown parameters by taking advantage of the big data and complex multi-physics simulation software. Future work will present the effect of spatial measurement density on the estimation accuracy and uncertainty reduction, and investigate prediction performances in a new experimental condition subject to the same sand packing (e.g.

[*Zhang et al.*, 2007; *Kokkinaki et al.*, 2013]) based on the estimated K field presented in this study.

Appendix A: Generalized Eigenvalue Problem

The Generalized Eigenvalue Problem (GEP) is defined as

$$\mathbf{Ax} = \lambda \mathbf{Bx} \tag{A1}$$

where, $\mathbf{B}$ is symmetric positive definite and $\mathbf{A}$ is symmetric. We can transform GEP into a typical eigenvalue problem. since $\mathbf{B}$ is positive definite, it has a Cholesky decomposition $\mathbf{B} = \mathbf{L}\mathbf{L}^\top$. Define $\mathbf{y} = \mathbf{L}^\top \mathbf{x}$ and multiplying both sides of (A1) by $\mathbf{L}^{-1}$, we have

$$\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-\top}\mathbf{L}^\top\mathbf{x} = \lambda\mathbf{L}^\top\mathbf{x} \qquad \Rightarrow \qquad \mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-\top}\mathbf{y} = \lambda\mathbf{y} \tag{A2}$$

which is an eigenvalue problem and hence, we can use any algorithm for eigenvalue problem to solve GEPs. However, computing the Cholesky decomposition is not computationally feasible in many cases and alternative methods can be found in [*Saad*, 2011; *Saibaba et al.*, 2015].

In our case, $\mathbf{A} := \mathbf{HQH}^\top$ and $\mathbf{B} := \mathbf{R}$

$$\mathbf{HQH}^\top\mathbf{x} = \lambda\mathbf{Rx} \tag{A3}$$

In many cases including the application considered in the paper, error is assumed to be an independent and identically distributed, i.e., $\mathbf{R} = \sigma^2\mathbf{I}$ and we can use eigenproblem solvers and the solution becomes

$$\mathbf{\Lambda} = \frac{1}{\sigma^2}\mathbf{\Lambda}_{\mathbf{HQH}^\top}, \quad \mathbf{U} = \frac{1}{\sigma}\mathbf{U}_{\mathbf{HQH}^\top} \tag{A4}$$

where the columns of $\mathbf{\Lambda}_{\mathbf{HQH}^\top}$ and the diagonal values of $\mathbf{U}_{\mathbf{HQH}^\top}$ are the eigenvalues and eigenvectors of $\mathbf{HQH}^\top$, respectively. For a general diagonal error matrix $\mathbf{R}$, the solution

is given by

$$\mathbf{\Lambda} = \mathbf{\Lambda}_{\mathbf{R}^{-\frac{1}{2}}\mathbf{HQH}^\top\mathbf{R}^{-\frac{1}{2}}} \quad \mathbf{U} = \mathbf{R}^{-\frac{1}{2}}\mathbf{U}_{\mathbf{R}^{-\frac{1}{2}}\mathbf{HQH}^\top\mathbf{R}^{-\frac{1}{2}}} \tag{A5}$$

where the columns of $\mathbf{\Lambda}_{\mathbf{R}^{-\frac{1}{2}}\mathbf{HQH}^\top\mathbf{R}^{-\frac{1}{2}}}$ and the diagonal values of $\mathbf{U}_{\mathbf{R}^{-\frac{1}{2}}\mathbf{HQH}^\top\mathbf{R}^{-\frac{1}{2}}}$ are the eigenvalues and eigenvectors of $\mathbf{R}^{-\frac{1}{2}}\mathbf{HQH}^\top\mathbf{R}^{-\frac{1}{2}}$, respectively.

Appendix B: Fast posterior variance computation

The diagonal entries of the posterior covariance matrix $\mathbf{V}$ in (6) are often presented as the estimation variance and can be computed without constructing $\mathbf{V}$ explicitly as

$$\mathbf{V}_{ii} = \mathbf{Q}_{ii} - \begin{bmatrix}\mathbf{HQ}_i \\ \mathbf{X}_i^\top\end{bmatrix}^\top \begin{bmatrix}\mathbf{HQH}^\top + \mathbf{R} & \mathbf{HX} \\ (\mathbf{HX})^\top & \mathbf{0}\end{bmatrix}^{-1} \begin{bmatrix}\mathbf{HQ}_i \\ \mathbf{X}_i^\top\end{bmatrix} \tag{B1}$$

where $\mathbf{V}_{ii}$ is the i-th diagonal element of $\mathbf{V}$, $\mathbf{Q}_{ii}$ is the i-th diagonal entry or the prior variance of i-th parameter, $\mathbf{HQ}_i$ is the i-th column of $\mathbf{HQ}$, and $\mathbf{X}_i^\top$ is the i-th column of $\mathbf{X}^\top$. Plugging (13) into (B1) yields

$$\begin{aligned}\mathbf{V}_{ii} = \mathbf{Q}_{ii} &- (\mathbf{HQ}_i)^\top\mathbf{\Psi}^{-1}\mathbf{HQ}_i + (\mathbf{HQ}_i)^\top\mathbf{\Psi}^{-1}\mathbf{\Phi}\mathbf{S}^{-1}\mathbf{\Phi}^\top\mathbf{\Psi}^{-1}\mathbf{HQ}_i \\ &- 2(\mathbf{HQ}_i)^\top\mathbf{\Psi}^{-1}\mathbf{\Phi}\mathbf{S}^{-1}\mathbf{X}_i^\top + \mathbf{X}_i\mathbf{S}\mathbf{X}_i^\top\end{aligned} \tag{B2}$$

However, repetitive multiplications of the 1 by m vector $(\mathbf{HQ}_i)^\top$ with m by m matrices (e.g., $\mathbf{\Psi}^{-1}$) for $i = 1, \cdots, m$ would be time consuming if m is large, i.e., $m = \mathcal{O}(10^6)$. A fast way to evaluate the estimation variance is to reduce the size of the repetitive matrix-vector multiplications by reformulating (B2). In PCGA, $\mathbf{HQ} \approx \mathbf{H}\boldsymbol{Z}_\kappa\boldsymbol{Z}_\kappa^\top \approx \sum_{i=1}^{\kappa}\boldsymbol{\eta}_i\boldsymbol{\zeta}_i^\top = \boldsymbol{\mathcal{H}}\boldsymbol{Z}_\kappa^\top$ where the n by κ matrix $\boldsymbol{\mathcal{H}}$ consists of column vectors $\boldsymbol{\eta}_{i=1,\cdots,\kappa}$, then

$$\begin{aligned}\mathbf{V}_{ii} = \mathbf{Q}_{ii} &- \boldsymbol{Z}_i\left(\boldsymbol{\mathcal{H}}^\top\mathbf{\Psi}^{-1}\boldsymbol{\mathcal{H}}\right)\boldsymbol{Z}_i^\top + \boldsymbol{Z}_i\left(\boldsymbol{\mathcal{H}}^\top\mathbf{\Psi}^{-1}\mathbf{\Phi}\mathbf{S}^{-1}\mathbf{\Phi}^\top\mathbf{\Psi}^{-1}\boldsymbol{\mathcal{H}}\right)\boldsymbol{Z}_i^\top \\ &- 2\boldsymbol{Z}_i\left(\boldsymbol{\mathcal{H}}^\top\mathbf{\Psi}^{-1}\mathbf{\Phi}\mathbf{S}^{-1}\right)\mathbf{X}_i^\top + \mathbf{X}_i\mathbf{S}\mathbf{X}_i^\top\end{aligned} \tag{B3}$$

where $\boldsymbol{Z}_i$ is the i-th "row" (a 1 by κ vector) of $\boldsymbol{Z}_\kappa$. In (B2), one has to evaluate products of a 1 by m vector and m by m matrices and the overall cost for the variance map $\mathbf{V}_{ii,i=1,\cdots,m}$ is $\mathcal{O}(m^3)$. In (B3), on the other hand, only products of a κ by 1 vector and κ by κ matrices are needed with the cost of $\mathcal{O}(m\kappa^2)$.

Acknowledgments. The research was funded by the National Science Foundation through its ReNUWIt Engineering Research Center (www.renuwit.org; NSF EEC-1028968). HY was supported as part of the Center for Frontiers of Subsurface Energy Security, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award Number DE-SC0001114. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. We thank Arvind Saibaba for his helpful comments on preconditioner construction and Amalia Kokkinaki for constructive suggestions.

References

Aravkin, A., M. P. Friedlander, F. J. Herrmann, and T. van Leeuwen (2012), Robust inversion, dimensionality reduction, and randomized sampling, *Mathematical Programming*, *134*(1), 101–125.

Barnhart, K., I. Urteaga, Q. Han, A. Jayasumana, and T. Illangasekare (2010), On integrating groundwater transport models with wireless sensor networks, *Ground Water*, *48*(5), 771–780.

606 Benzi, M., G. H. Golub, J. Liesen, ouml, and rg (2005), Numerical solution of saddle point
607 problems, *Acta Numerica*, *14*, 1–137.
608 Brown, P. N., and Y. Saad (1990), Hybrid krylov methods for nonlinear-systems of equa-
609 tions, *Siam Journal on Scientific and Statistical Computing*, *11*(3), 450–481.
610 Carrera, J., A. Alcolea, A. Medina, J. Hidalgo, and L. J. Slooten (2005), Inverse problem
611 in hydrogeology, *Hydrogeology Journal*, *13*(1), 206–222.
612 Cirpka, O. A., and P. K. Kitanidis (2000), Sensitivity of temporal moments calculated
613 by the adjoint-state method and joint inversing of head and tracer data, *Advances in*
614 *Water Resources*, *24*(1), 89–103.
615 Cui, T., J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini (2014), Likelihood-
616 informed dimension reduction for nonlinear inverse problems, *Inverse Problems*, *30*(11).
617 Deutsch, C. V., and A. G. Journel (1998), *GSLIB, geostatistical software library and user's*
618 *guide*, Applied geostatistics series, 2nd ed., Oxford University Press, New York.
619 Doherty, J. E., and R. J. Hunt (2010), *Approaches to highly parameterized inversion: a*
620 *guide to using PEST for groundwater-model calibration*, US Department of the Interior,
621 US Geological Survey.
622 Epanomeritakis, I., V. Akcelik, O. Ghattas, and J. Bielak (2008), A newton-cg method for
623 large-scale three-dimensional elastic full-waveform seismic inversion, *Inverse Problems*,
624 *24*(3).
625 Fakhreddine, S., J. Lee, P. K. Kitanidis, S. Fendorf, and M. Rolle (2015), Imaging geo-
626 chemical heterogeneities using inverse reactive transport modeling: An example relevant
627 for characterizing arsenic mobilization and distribution advances in water resources, *Ad-*
628 *vances in Water Resources, in review.*

Flath, H. P., L. C. Wilcox, V. Akcelik, J. Hill, B. V. Waanders, and O. Ghattas (2011), Fast algorithms for bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial hessian approximations, *Siam Journal on Scientific Computing*, *33*(1), 407–432.

Frauenfelder, P., C. Schwab, and R. A. Todor (2005), Finite elements for elliptic problems with stochastic coefficients, *Computer Methods in Applied Mechanics and Engineering*, *194*(2-5), 205–228.

Haber, E., and U. M. Ascher (2001), Preconditioned all-at-once methods for large, sparse parameter estimation problems, *Inverse Problems*, *17*(6), 1847–1864.

Haber, E., M. Chung, and F. Herrmann (2012), An effective method for parameter estimation with pde constraints with multiple right-hand sides, *Siam Journal on Optimization*, *22*(3), 739–757.

Hampson, G., J. Stefani, and F. Herkenhoff (2008), Acquisition using simultaneous sources, *The Leading Edge*, *27*(7), 918–923.

Harbaugh, A. W., E. R. Banta, M. C. Hill, and M. G. McDonald (2000), Modflow-2000, the us geological survey modular ground-water model: User guide to modularization concepts and the ground-water flow process, *Tech. Rep. 2000-92*, Geological Survey (U.S.).

Harvey, C. F., and S. M. Gorelick (1995a), Mapping hydraulic conductivity - sequential conditioning with measurements of solute arrival time, hydraulic-head, and local conductivity, *Water Resources Research*, *31*(7), 1615–1626.

Harvey, C. F., and S. M. Gorelick (1995b), Temporal moment-generating equations - modeling transport and mass-transfer in heterogeneous aquifers, *Water Resources Research*,

652 *31*(8), 1895–1911.

653 Hochstetler, D. L., W. Barrash, C. Leven, M. Cardiff, F. Chidichimo, and P. K. Kitanidis
654 (2015), Hydraulic tomography: Continuity and discontinuity of high-k and low-k zones,
655 *Groundwater*, pp. n/a–n/a.

656 Jose, S. C., M. A. Rahman, and O. A. Cirpka (2004), Large-scale sandbox experiment
657 on longitudinal effective dispersion in heterogeneous porous media, *Water Resources*
658 *Research*, *40*(12).

659 Kaipio, J., and E. Somersalo (2007), Statistical inverse problems: Discretization, model re-
660 duction and inverse crimes, *Journal of Computational and Applied Mathematics*, *198*(2),
661 493–504.

662 Kitanidis, P. K. (1983), Statistical estimation of polynomial generalized covariance func-
663 tions and hydrologic applications, *Water Resources Research*, *19*(4), 909–921.

664 Kitanidis, P. K. (1991), Orthonormal residuals in geostatistics - model criticism and
665 parameter-estimation, *Mathematical Geology*, *23*(5), 741–758.

666 Kitanidis, P. K. (1993), Generalized covariance functions in estimation, *Mathematical*
667 *Geology*, *25*(5), 525–540.

668 Kitanidis, P. K. (1995), Quasi-linear geostatistical theory for inversing, *Water Resources*
669 *Research*, *31*(10), 2411–2419.

670 Kitanidis, P. K. (2010), Bayesian and geostatistical approaches to inverse problems, in
671 *Large-Scale Inverse Problems and Quantification of Uncertainty*, pp. 71–85, John Wiley
672 & Sons, Ltd.

673 Kitanidis, P. K., and J. Lee (2014), Principal component geostatistical approach for large-
674 dimensional inverse problems, *Water Resources Research*, *50*(7), 5428–5443.

675 Kokkinaki, A., D. M. O'Carroll, C. J. Werth, and B. E. Sleep (2013), Coupled simulation of
676 dnapl infiltration and dissolution in three-dimensional heterogeneous domains: Process
677 model validation, *Water Resources Research*, *49*(10), 7023–7036.

678 Krebs, J. R., J. E. Anderson, D. Hinkley, R. Neelamani, S. Lee, A. Baumstein, and M. D.
679 Lacasse (2009), Fast full-wavefield seismic inversion using encoded sources, *Geophysics*,
680 *74*(6), Wcc177–Wcc188.

681 Lee, J., and P. K. Kitanidis (2014), Large-scale hydraulic tomography and joint inversion
682 of head and tracer data using the principal component geostatistical approach (pcga),
683 *Water Resources Research*, *50*(7), 5410–5427.

684 Lee, J., A. Kokkinaki, Y. Li, and P. K. Kitandis (2015), Fast large-scale inversion for
685 deep aquifer characterization, in *TOUGH2 Symposium*, Lawrence Berkeley National
686 Laboratory, Berkeley, California.

687 McLaughlin, D., and L. R. Townley (1996), A reassessment of the groundwater inverse
688 problem, *Water Resources Research*, *32*(5), 1131–1161.

689 Nowak, W., and O. A. Cirpka (2006), Geostatistical inference of hydraulic conductivity
690 and dispersivities from hydraulic heads and tracer data, *Water Resources Research*,
691 *42*(8), W08,416.

692 Oliver, D. S., A. C. Reynolds, and N. Liu (2008), *Inverse theory for petroleum reservoir*
693 *characterization and history matching*, Cambridge University Press, Cambridge ; New
694 York.

695 Orellana, G., and D. Haigh (2008), New trends in fiber-optic chemical and biological
696 sensors, *Current Analytical Chemistry*, *4*(4), 273–295.

Paige, C. C., and M. A. Saunders (1975), Solution of sparse indefinite systems of linear equations, *Siam Journal on Numerical Analysis*, *12*(4), 617–629.

Pamukcu, S., and E. Ghazanfari (2014), Geosensing for developing sustainable responses to environmental hazards underground, in *Geo-Congress 2014 Keynote Lectures*, pp. 117–139.

Saad, Y. (2011), Numerical methods for large eigenvalue problems, Society for Industrial and Applied Mathematics,.

Saad, Y., and M. H. Schultz (1986), Gmres - a generalized minimal residual algorithm for solving nonsymmetric linear-systems, *Siam Journal on Scientific and Statistical Computing*, *7*(3), 856–869.

Saibaba, A. K., and P. K. Kitanidis (2015), Fast computation of uncertainty quantification measures in the geostatistical approach to solve inverse problems, *Advances in Water Resources*, *82*, 124–138.

Saibaba, A. K., J. Lee, and P. K. Kitanidis (2015), Randomized square-root free algorithms for generalized hermitian eigenvalue problems, *Numerical Linear Algebra with Applications, in press.*

Smith, R. C. (2014), *Uncertainty quantification : theory, implementation, and applications*, Computational science & engineering, SIAM, Society for Industrial and Applied Mathematics, Philadelphia.

Stuart, A. M. (2010), Inverse problems: A bayesian perspective, *Acta Numerica 2010, Vol 19*, *19*, 451–559.

Valocchi, A. J. (1986), Effect of radial flow on deviations from local equilibrium during sorbing solute transport through homogeneous soils, *Water Resources Research*, *22*(12),

720 1693–1701.

721 Yin, D. T., and W. A. Illman (2009), Hydraulic tomography using temporal moments of
722 drawdown recovery data: A laboratory sandbox study, *Water Resources Research*, *45*.

723 Yoon, H., and S. A. McKenna (2012), Highly parameterized inverse estimation of hydraulic
724 conductivity and porosity in a three-dimensional, heterogeneous transport experiment,
725 *Water Resources Research*, *48*.

726 Yoon, H., C. Y. Zhang, C. J. Werth, A. J. Valocchi, and A. G. Webb (2008), Numerical
727 simulation of water flow in three dimensional heterogeneous porous media observed in
728 a magnetic resonance imaging experiment, *Water Resources Research*, *44*(6).

729 Zhang, C. Y., C. J. Werth, and A. G. Webb (2007), Characterization of napl source zone
730 architecture and dissolution kinetics in heterogeneous porous media using magnetic
731 resonance imaging, *Environmental Science & Technology*, *41*(10), 3672–3678.

732 Zheng, C., and P. P. Wang (1999), MT3DMS: A modular three-dimensional multispecies
733 transport model for simulation of advection, dispersion, and chemical reactions of
734 contaminants in groundwater systems; documentation and user's guide, *Tech. Rep.*
735 *SERDP-99-1*, DTIC Document, Vicksburg, Mississippi.

736 Zhu, J. F., and T. C. J. Yeh (2006), Analysis of hydraulic tomography using temporal
737 moments of drawdown recovery data, *Water Resources Research*, *42*(2).

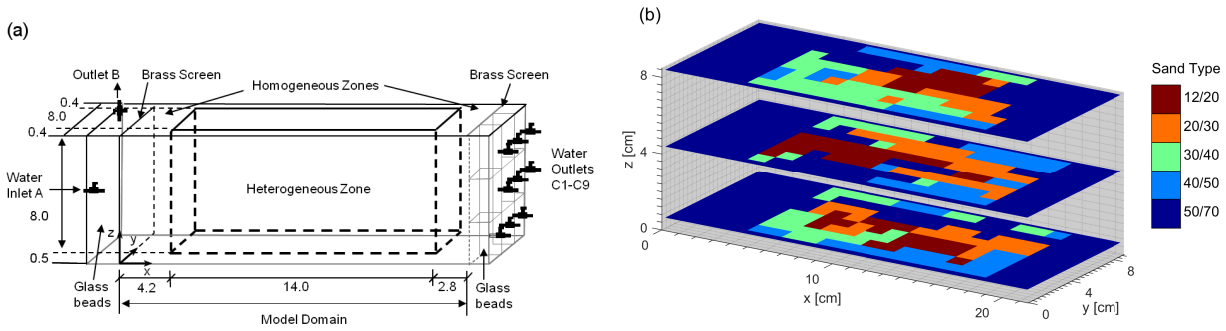

Figure 1. (a) Illustration of 3-D flowcell [*Yoon et al.*, 2008] and (b) hydraulic conductivity distribution in 3 layers (out of 8 layers in total).

Table 1. Properties of Sands

Sand Type	Measured K (cm/min)	Mean Grain Size (cm)
12/20	25.08	0.11
20/30	13.44	0.072
30/40	6.72	0.053
40/50	3.78	0.036
50/70	2.03	0.026

Table 2. Simulation and Inversion Parameters for the Synthetic Cases

Parameter	Description	Value	
		Case 1	Case 2
	Simulation Parameters		
L_x, L_y, L_z	domain length and width (cm)	21, 9, 8.5	
$\Delta x, \Delta y, \Delta z$	grid spacing (cm)	0.25	
θ	porosity	0.355	
α_l	longitudinal dispersivity (cm)	$0.013 \sim 0.055$ [c]	
α_t	transverse dispersivity (cm)	$0.1 \times \alpha_l$	
	True K field Generation		
	$\ln K$ generation method	Gaussian[a]	SISIM[b]
	Measurement Error		
n_C	number of concentration measurement	5,777,408	
σ_C	standard deviation of measurement error (-)	0.1	
	Inversion Parameters		
m	the number of unknowns	57,344	99,072
n_{obs}	number of travel time measurements (min)	51,584	
$q(x, x')$	covariance kernel	$q(x, x') = \sigma^2_{\ln K} \exp(-\lvert x - x' \rvert / l)$	
$\sigma^2_{\ln K}$	prior variance (cm/min^2)	0.5	0.1
l_x, l_y, l_z	scale parameter l in x, y, z (cm)	4, 2, 2	2, 1, 1
σ_t	standard deviation of measurement error for travel time (min)	3	
δ	finite difference interval for PCGA	0.005	

[a] exponential covariance with $\sigma^2_{\ln K} = 0.5$ (cm^2) and $l_x, l_y, l_z = 4, 2, 2$ (cm)

[b] *Zhang et al.* [2007]

[c] The dispersivity field was determined based on the mean grain size of the sands used in the experiment; The detailed information can be found in *Yoon et al.* [2008]

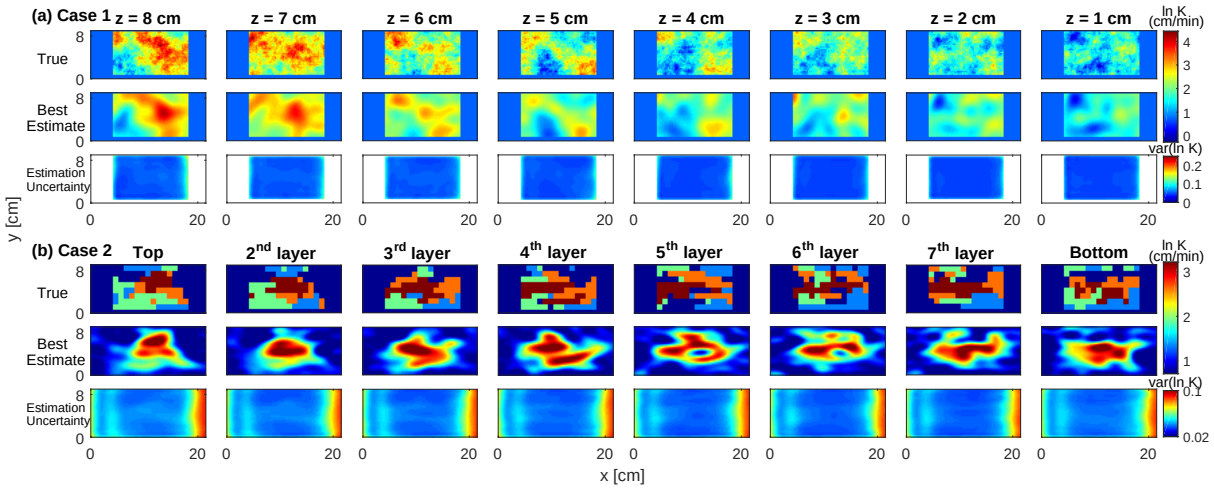

Figure 2. The true (first row), best estimate (second row) and estimation uncertainty (third row) for (a) Case 1 and (b) Case 2: (a) values at a specific height z are plotted; (b) averaged values over the depth of 1 cm are plotted in order to compare with the true field.

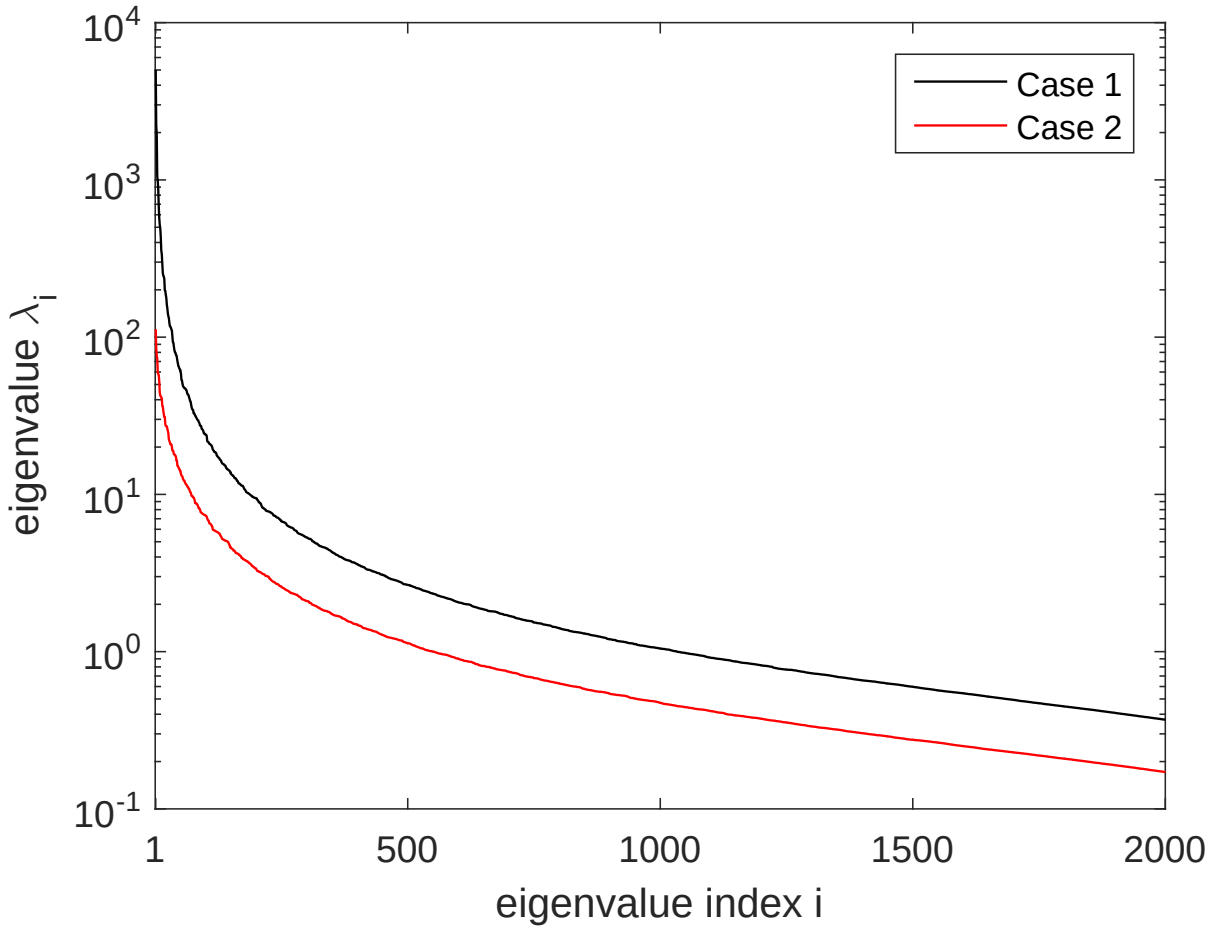

Figure 3. The eigenvalue spectrum of the prior covariance for Case 1 (black) and Case 2 (red).

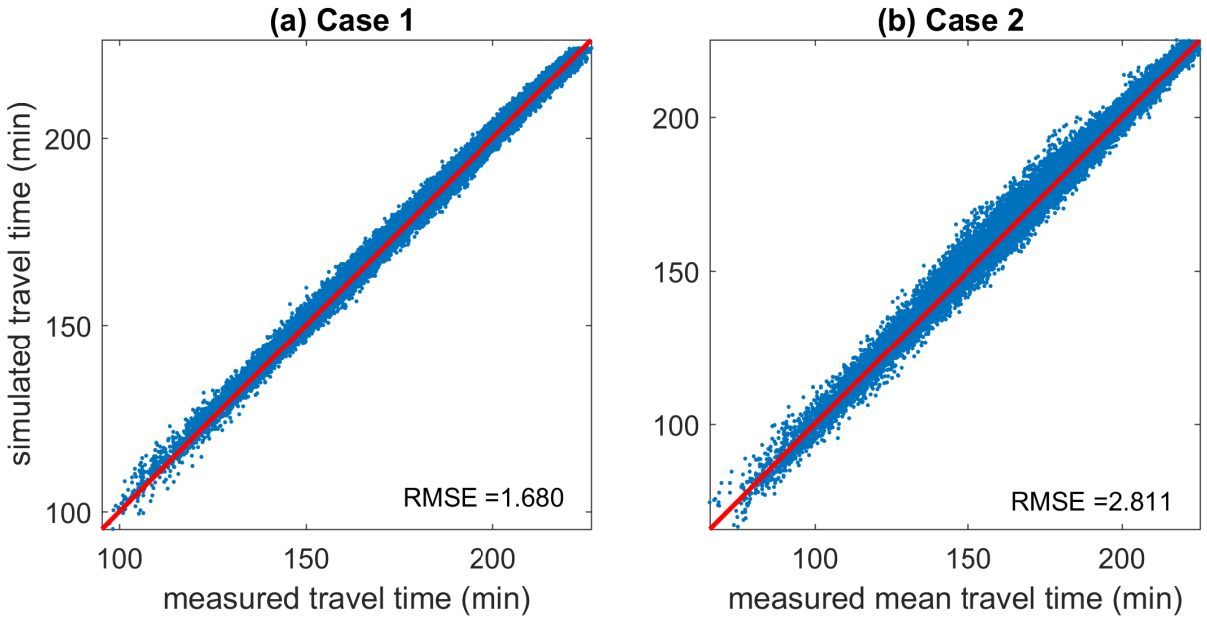

Figure 4. Measurement data fitting: measured versus simulated mean travel times from the best estimate for (a) Case 1 using $\kappa = 300$ and (b) Case 2 using $\kappa = 500$.

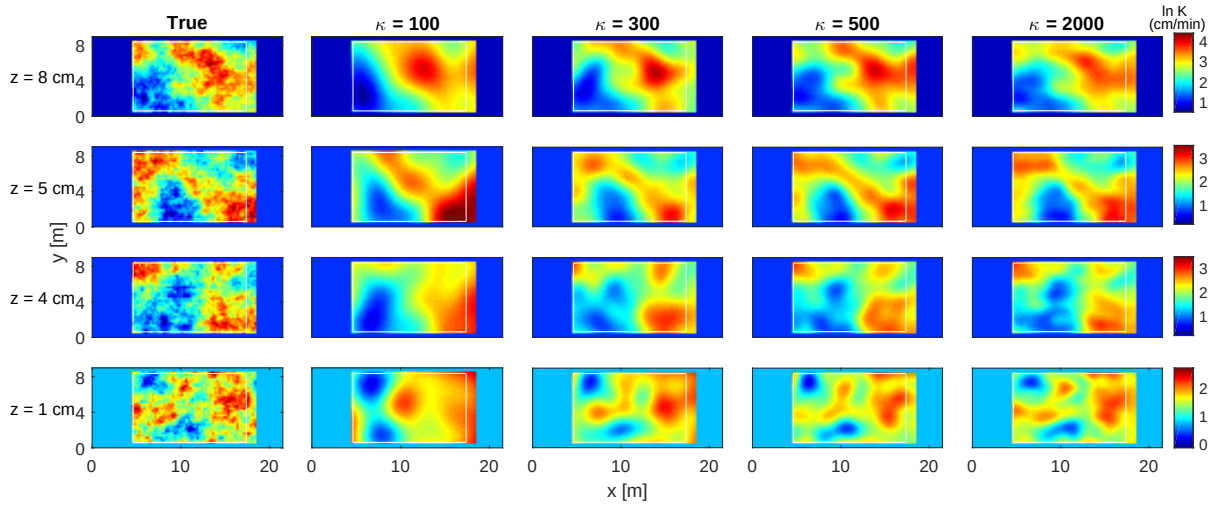

Figure 5. The best estimates with $\kappa = 100$, 300, 500, and 2,000 for Case 1.

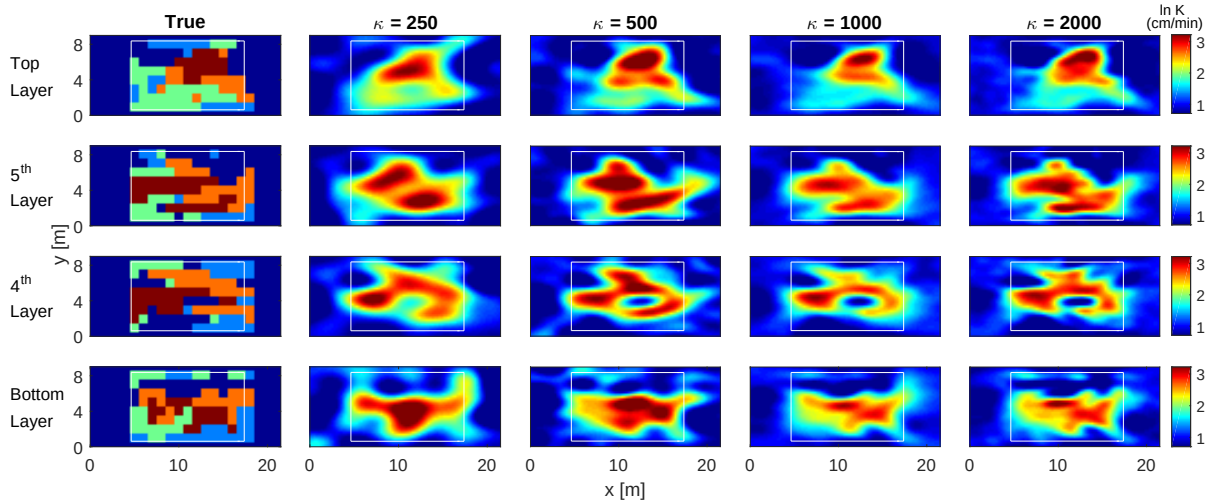

Figure 6. The best estimates with $\kappa = 250$, 500, 1000 and 2,000 for Case 2.

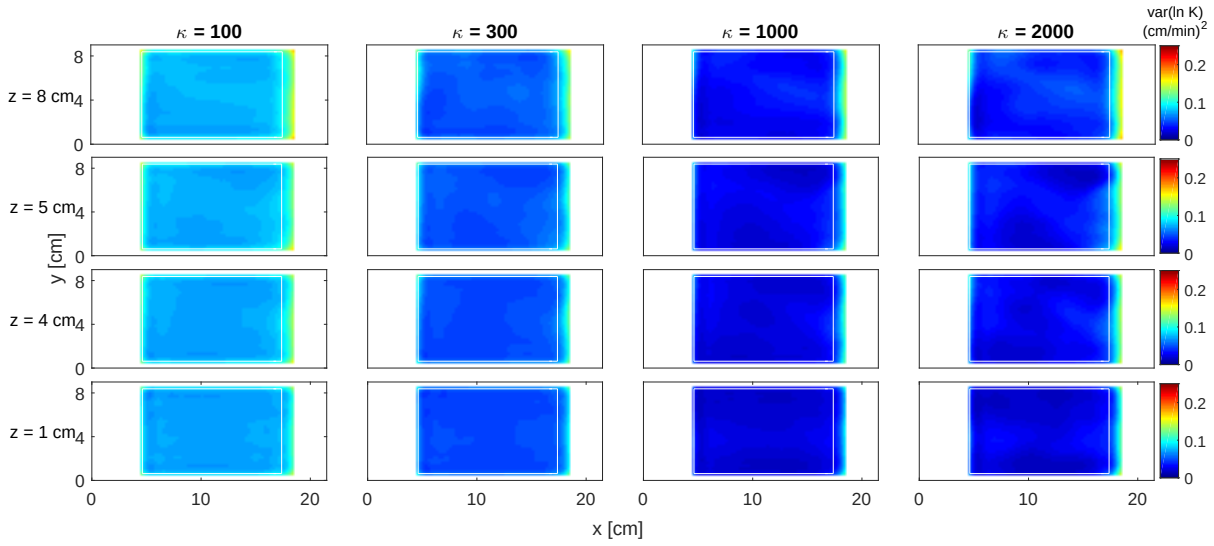

Figure 7. The estimation variance with $\kappa = 100$, 300, 500, and 2,000 for Case 1.

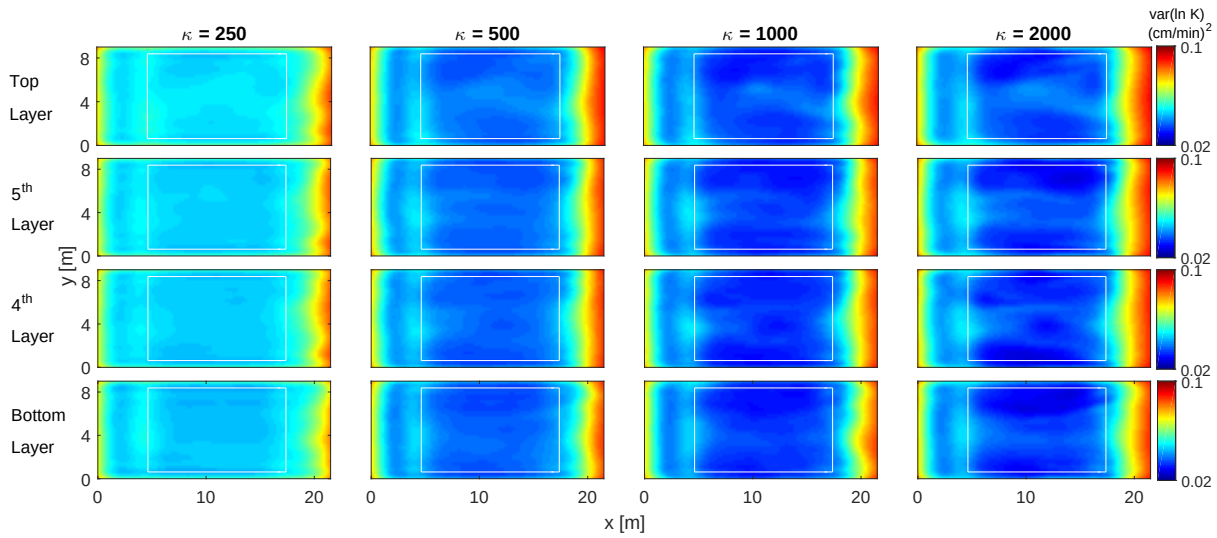

Figure 8. The estimation variance with $\kappa = 250$, 500, 1000, and 2,000 for Case 2.

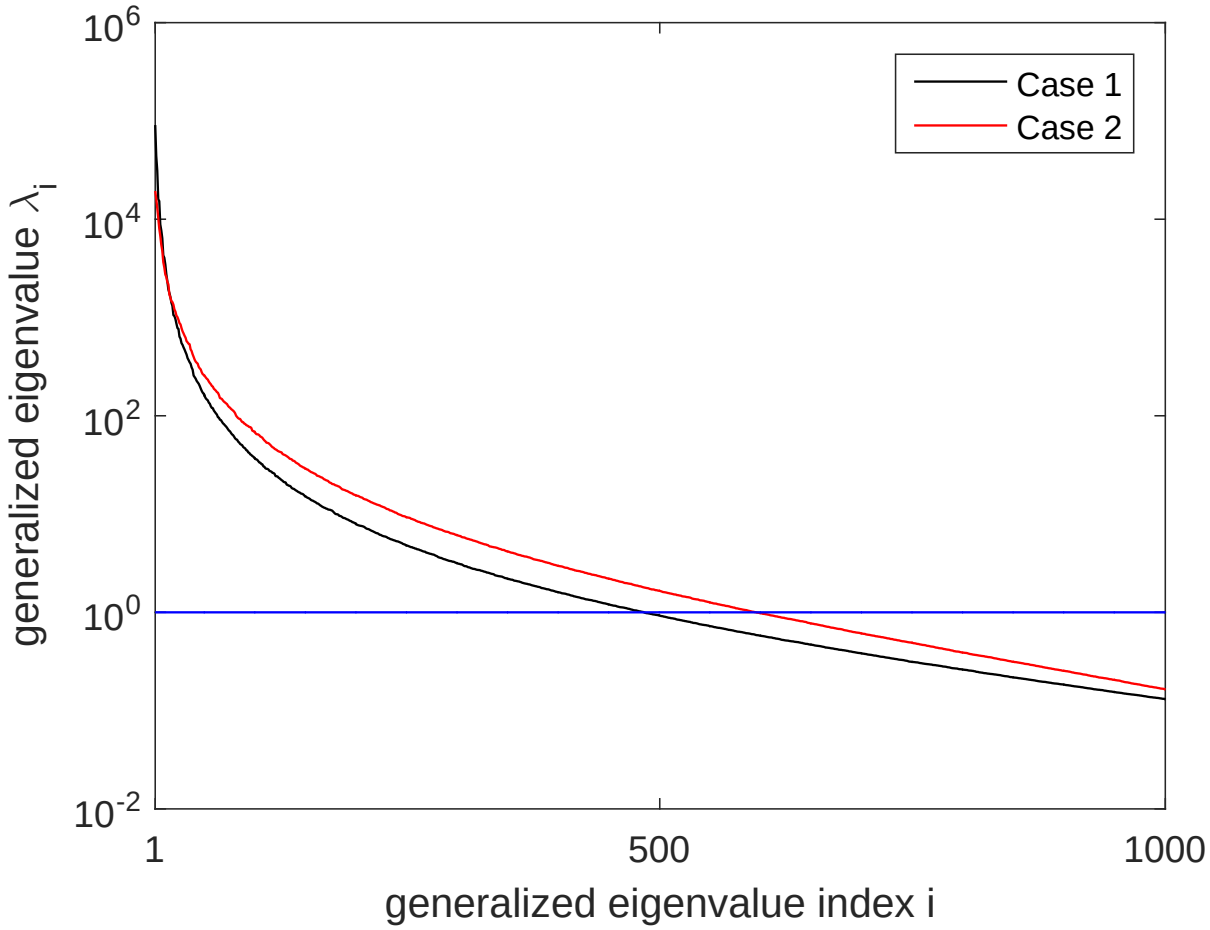

Figure 9. The generalized eigenvalue spectrum of $\mathbf{HQH}^\top$ and $\mathbf{R}$; $\kappa \leq 1$ for PCGA would result in the negligible error.

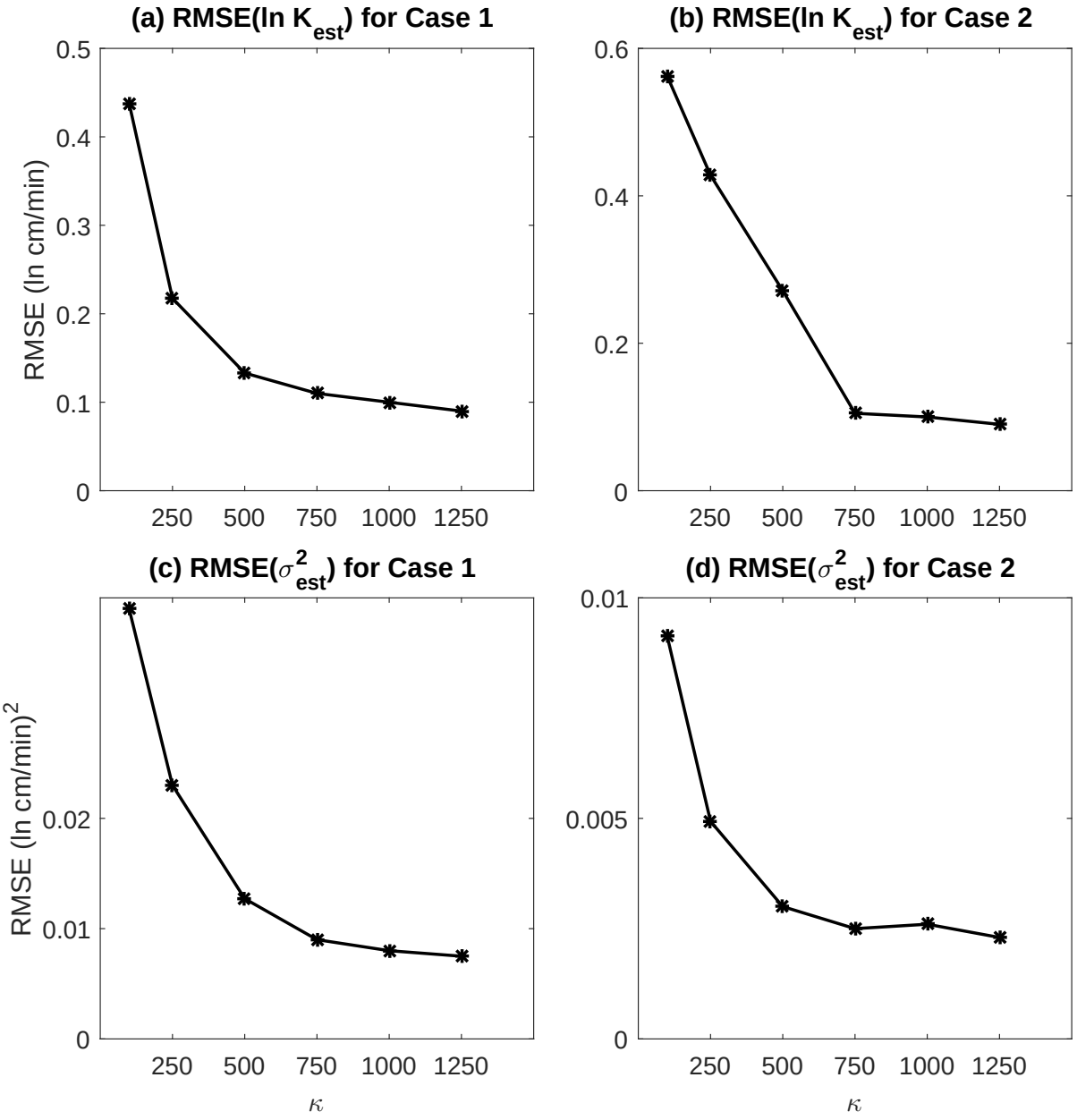

Figure 10. The root-mean-square error of the estimates for (a) Case 1 and (b) Case 2, and the estimation variance for (c) Case 1 and (d) Case 2.

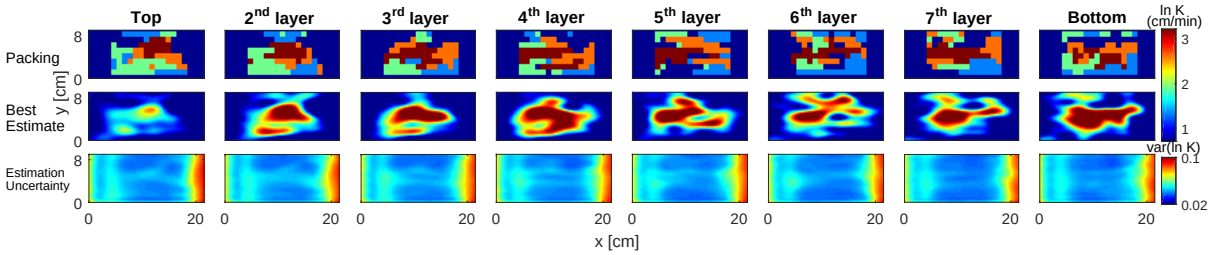

Figure 11. The packing pattern (first row), best estimate (second row) and estimation uncertainty (third row) for actual MRI travel time data inversion; averaged values over the depth of 1 cm are plotted in order to compare with the true packing pattern.

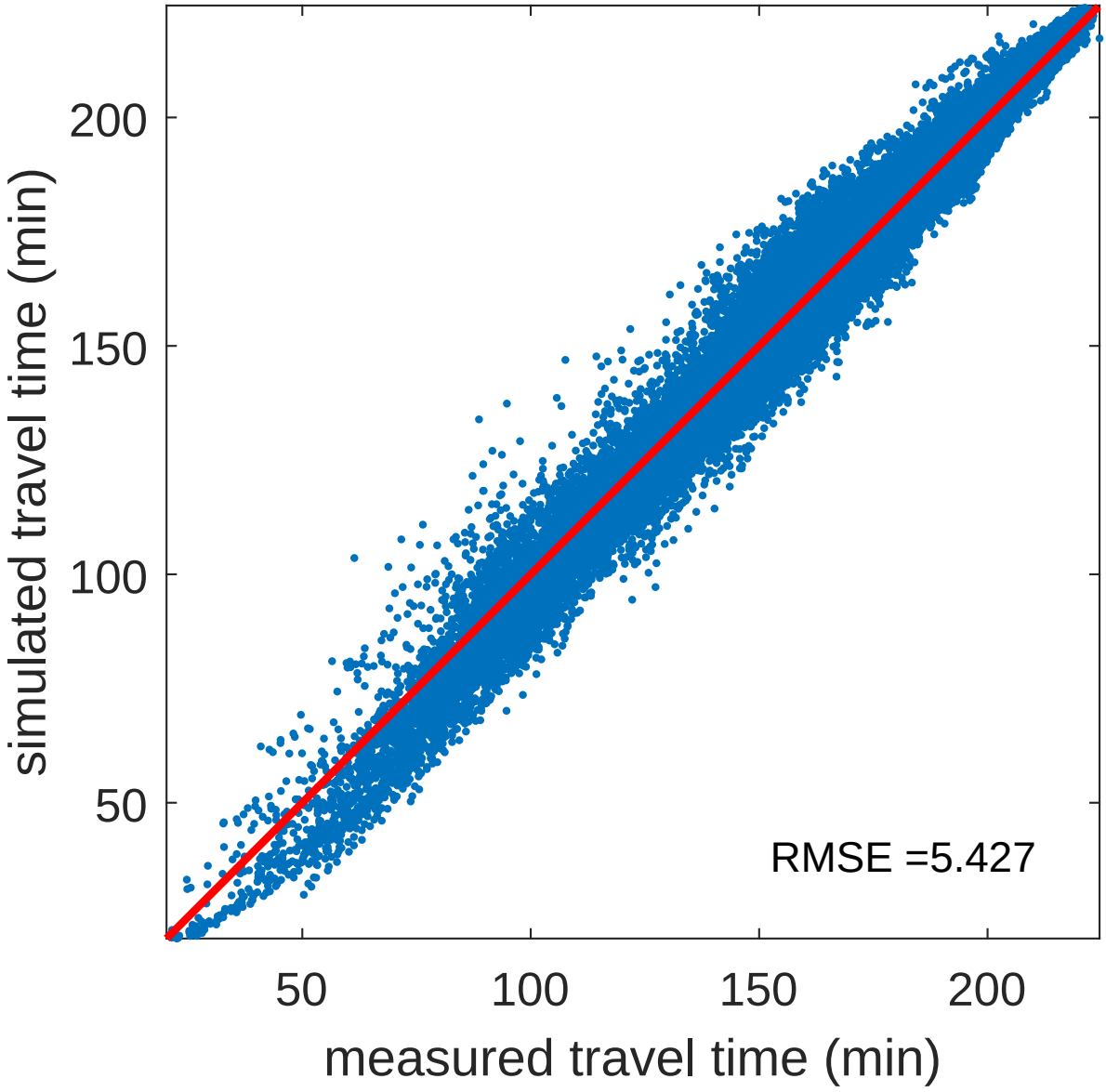

Figure 12. Measurement data fitting: measured versus simulated mean travel times from the best estimate.

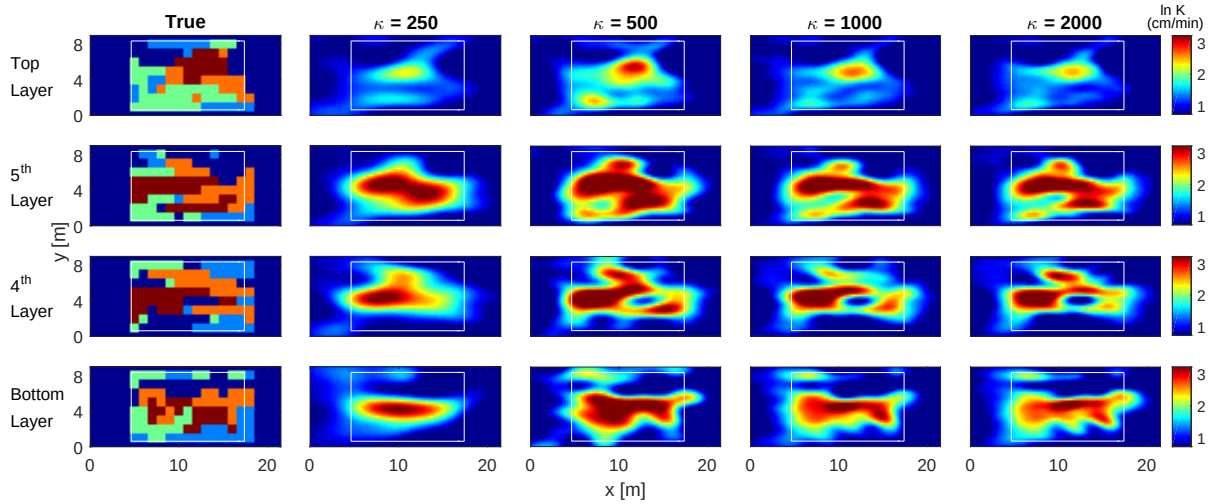

Figure 13. The best estimates with $\kappa = 250$, 500, 1000 and 2,000 for actual MRI data inversion.

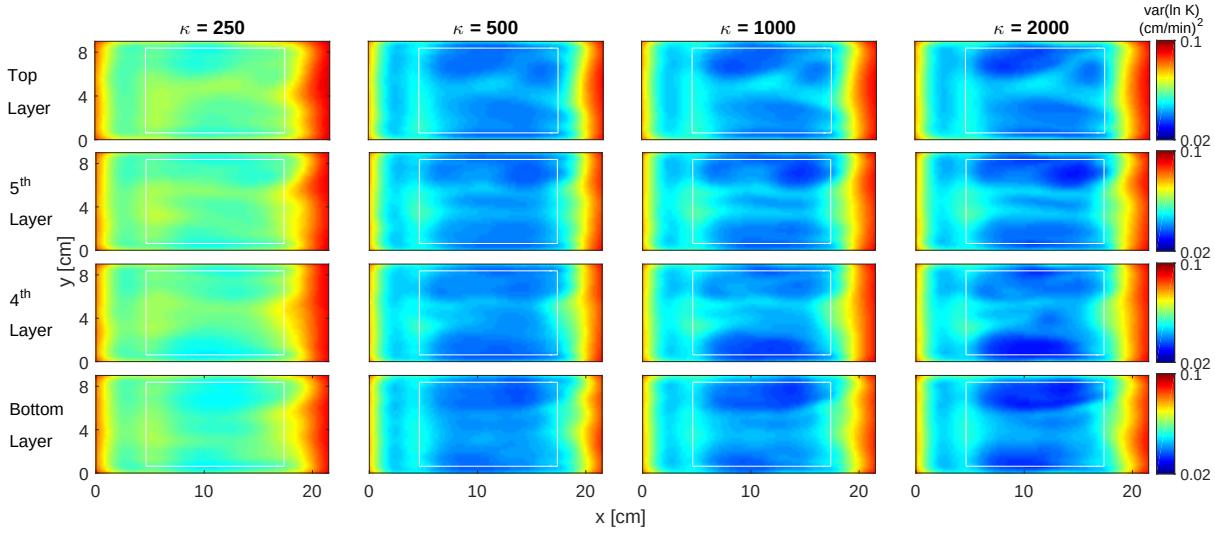

Figure 14. The estimation variance with $\kappa = 250$, 500, 1000, and 2,000 for actual MRI data inversion.

Table 3. Mapping accuracy of K distribution compared to the design packing pattern

Sand Type	Mapping Accuracy (%)		
	PCGA ($\kappa = 500$)	PEST [a]	PEST (with zonal information) [b]
12/20	97.3	64.5	86.9
20/30	98.0	60.8	84.8
30/40	89.4	48.5	81.1
40/50	72.6	78.2	91.2
50/70	70.8	57.1	82.3

[a] K_{IND} from *Yoon and McKenna* [2012]; K and porosity are estimated independently yielding the best mapping accuracy result except the case using the zonal information

[b] K_{Zone} from *Yoon and McKenna* [2012]; K field was parameterized by zonal boundaries from the design packing pattern