SAND2016-1781C

Scalable Subsurface Inverse Modeling Using Big Environmental Data Sets

Jonghyun Harry Lee[1], Hongkyu Yoon[2], Peter K. Kitanidis[1]

[1]Civil and Environmental Engineering, Stanford University [2]Geoscience Research and Applications Group, Sandia National Laboratories

Overview

With recent advances in sensor and computation technology, unprecedented large volumes of hydro-geophysical and geochemical data sets can be obtained to achieve high-resolution images of subsurface properties for more accurate and reliable subsurface flow and reactive transport prediction. However, massive data set inversion is challenging due to the cost of:

1) numerous simulation model runs to compute Jacobian matrix

2) large and dense matrix multiplications and storage

To tackle these challenges, the **Principal Component Geostatistical Approach [1-3]** has been proposed with following advantages:

- **Jacobian-free**: no need to compute/store full Jacobian matrix
- **forward model runs independent of the problem size**: often runs much smaller number of simulations in practice
- **linear scalability**: matrix computation/storage costs grow linearly with respect to the problem size
- **easy to implement**: linked with any **"black-box"** multi-physics simulation models **without invasive changes**
- **independent forward model executions in parallel**

Principal Component Geostatistical Approach

With $\mathbf{m}$ unknowns, $\mathbf{n_{obs}}$ measurements and forward model(s) $\mathbf{h}$, one needs to compute:

- Jacobian matrix $\mathbf{J}$, *i.e.*, sensitivity of the data to unknown parameters
- Jacobian products with the prior covariance $\mathbf{\Gamma}_{\text{prior}}$, *i.e.*, $\mathbf{J\Gamma}_{\text{prior}}$ and $\mathbf{J\Gamma}_{\text{prior}}\mathbf{J}^{\mathsf{T}}$

For large-scale/joint inversions (large $\mathbf{m}$ and $\mathbf{n_{obs}}$), one faces challenges

- time-consuming, invasive changes in multi-physics simulation code for efficient adjoint-state method implementation to evaluate Jacobian $\mathbf{J}$
- expensive Jacobian construction requiring $\mathbf{n_{obs}}$ ($\geq \mathcal{O}(10^4)$) simulations
- prohibitive large dense matrix multiplication/storage for $\mathbf{m}$ ($\geq \mathcal{O}(10^6)$)

In order to tackle these challenges, we developed PCGA that avoids expensive Jacobian evaluation and its matrix products (cross-covariance) by using a **fast truncated decomposition** [1] of the prior covariance

$$\mathbf{\Gamma}_{\text{prior}} \approx \mathbf{\Gamma}_{\kappa} = \Sigma_{i=1}^{\kappa}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^{\mathsf{T}} \quad \Leftarrow \textbf{scales linearly} \ (\text{Cost } \mathcal{O}(m\kappa^2))$$

and **finite-difference approximation**:

$$\mathbf{J}\boldsymbol{\zeta}_i \approx \frac{1}{\delta}\left[\mathbf{h}\left(\mathbf{s}+\delta\boldsymbol{\zeta}_i\right)-\mathbf{h}(\mathbf{s})\right], \quad \mathbf{J\Gamma}_{\text{prior}} \approx \Sigma_{i=1}^{\kappa}\left(\mathbf{J}\boldsymbol{\zeta}_i\right)\boldsymbol{\zeta}_i^{\mathsf{T}} \quad \Leftarrow \textbf{total } \kappa+1 \textbf{ simulations!}$$

Thus, PCGA can achieve **a significant speed-up with reasonable accuracy**, using simulation outputs **without modifying multi-physics simulation code**.

Application to conductivity estimation in a sand box from a massive MRI dataset

Objective: **reconstruct 99,072** $\log K$ **of a laboratory-scale sand box from 5,777,408 tracer concentration measurements obtained from MRI** [3].

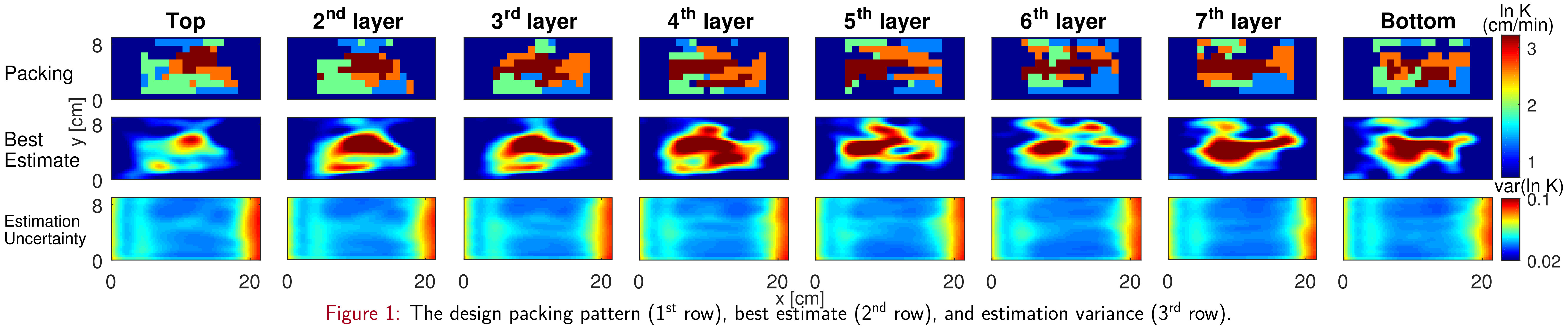

Figure 1: The design packing pattern (1st row), best estimate (2nd row), and estimation variance (3rd row).

Experimental & Numerical Setup

The entire flowcell has dimensions of 21.5 × 9 × 8.5 cm, and is packed with 1 cm cubes of five different sand types [4]:

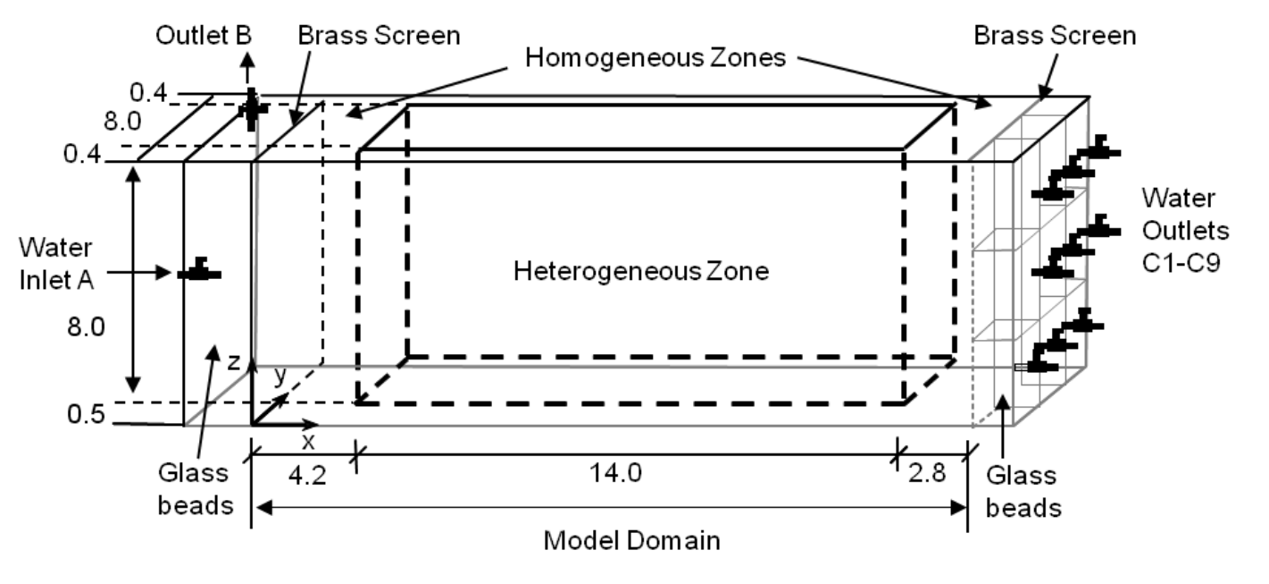

Figure 2: Illustration of 3-D flowcell

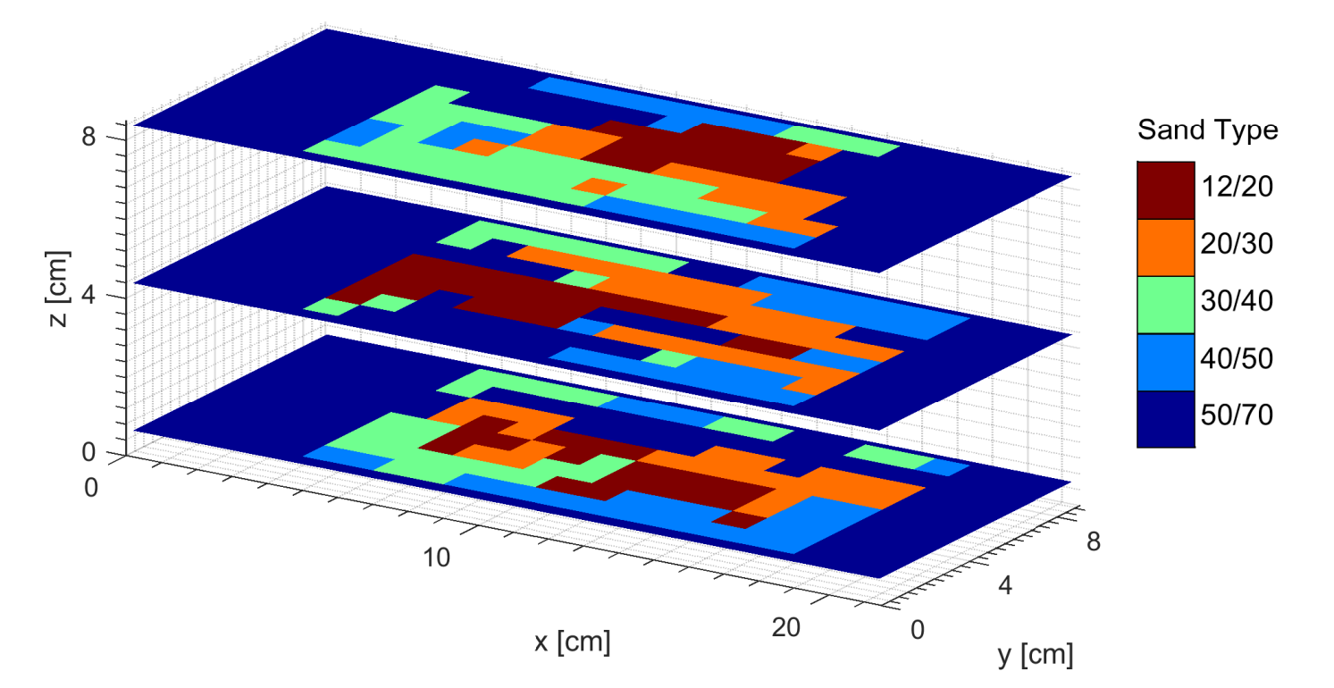

Figure 3: Sand packing in 3 layers (out of 8)

- Sand distribution was created using SISIM in GSLIB to construct a heterogeneous K field for the central portion (14 × 8 × 8 cm^3).
- Constant water flow rate with a uniform tracer concentration
- ∼ 6 million transient tracer concentrations were imaged using MRI at a resolution of 0.25^3 cm^3 at a regular interval time over the central region.
- A uniform grid spacing of 0.25 cm used for MODFLOW and MT3DMS.
- Forward simulations and inversions executed on a 36 core workstation.
- Using the zero-th temporal moment, 5,777,408 data records were reduced to 51,584 mean travel time records for the inversion.

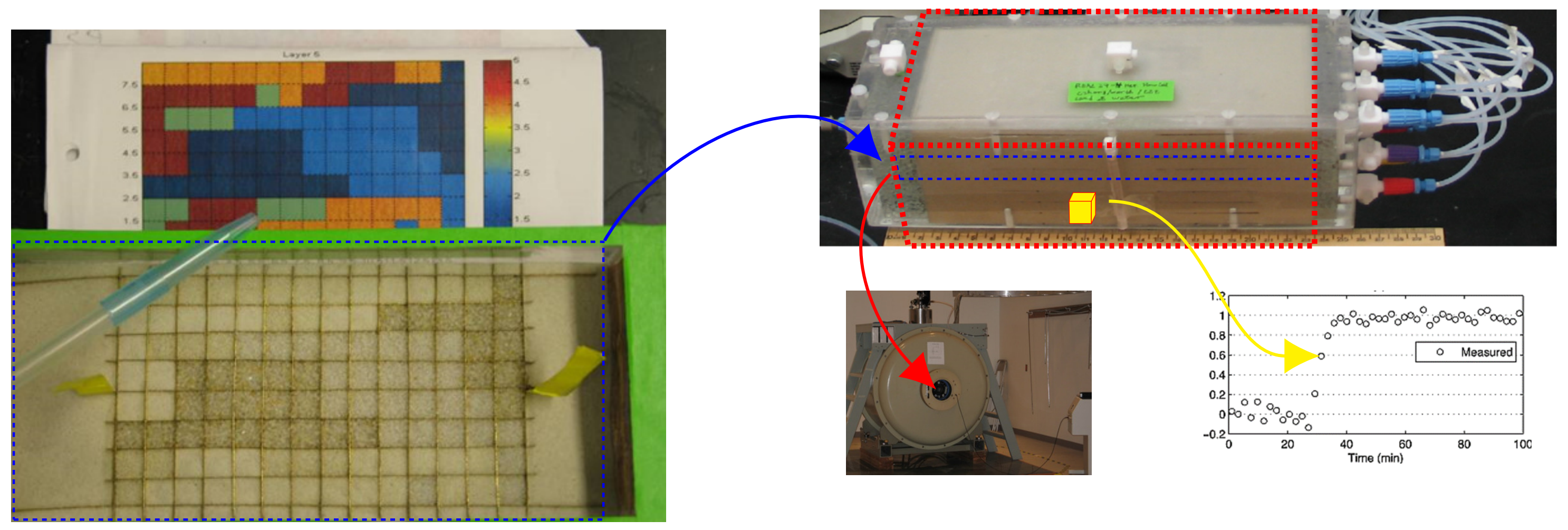

Figure 4: Packing with a brass divider with 1cm^3 openings (left), flowcell (upper right), MRI magnet (lower left) and normalized signal intensity at a voxel (lower right).

Results

- A total of **1,952** MODFLOW-MT3DMS simulation runs in **5 hours**.
- κ = 500 forward simulations for each iteration were enough to approximate a full inverse solution, which would require 51,584 runs.

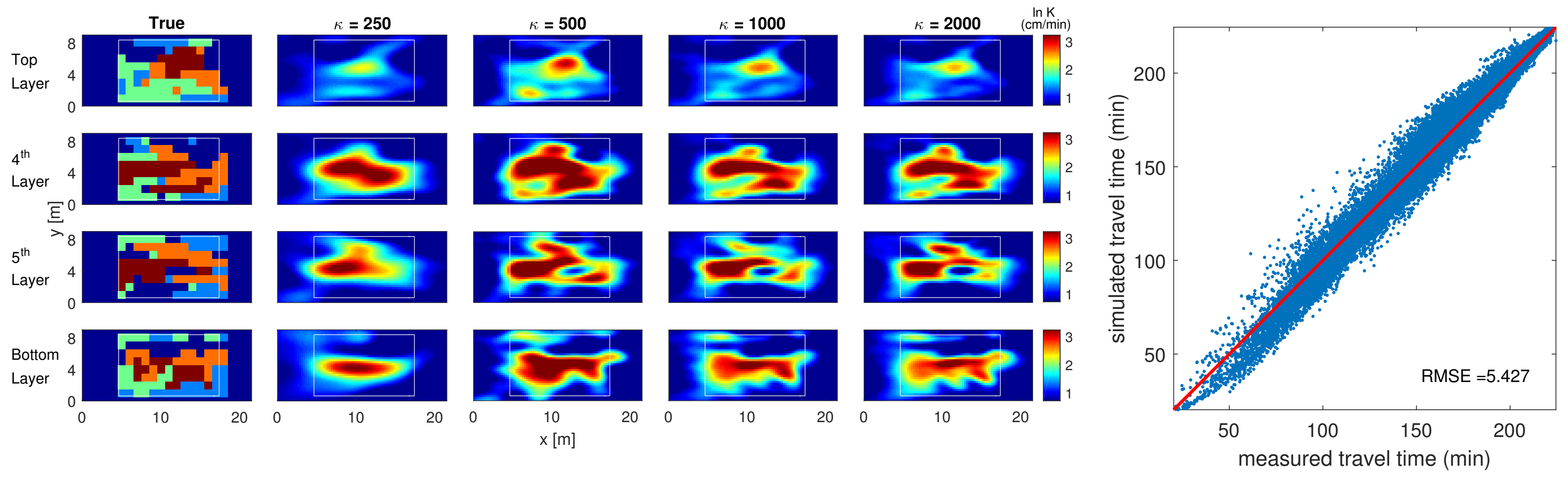

Figure 5: The best estimate with different κ values (left); data fitting with κ = 500 (right).

Conclusion

- PCGA performed a data-intensive inversion efficiently.
- Key patterns of the original sand packing design were identified.

References

[1] Lee and Kitanidis, 2014

[2] Kitanidis and Lee, 2014

[3] Lee *et. al.*, Scalable inverse modeling of huge datasets, *in review*

[4] Yoon *et. al.*, *WRR*, 2008

For more info:
stanford.edu/~jonghyun

Acknowledgment

This work is funded by the NSF 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.