Basis-Constrained Bayesian-McMC: Hydrologic Process Parameterization of Stochastic Geoelectrical Imaging of Solute Plumes

Erasmus Kofi Oware[1], Michael Awatey[1], Thomas Hermans[2], James Irving[3]
[1]*Department of Geology, SUNY at Buffalo, 126 Cooke Hall, Buffalo, NY 14260.*
[2]*Department of Geology, Ghent University, Belgium.*
[3]*Applied and Environmental Geophysics Group, Institute of Earth Sciences, University of Lausanne, Switzerland.*

Summary

Bayesian Markov-chain-Monte-Carlo (McMC) techniques are widely used in geophysics due to their ability to estimate multiple outcomes that enable uncertainty assessment. Standard McMC sampling methods, however, can become computationally intractable for high-dimensional problems. We present a Bayesian McMC framework that re-parameterizes the McMC procedure in terms of an optimal basis, thereby enabling a sparse representation of the desired features leading to dimensionality reduction. The approach also implicitly incorporates the physics of the underlying process. We demonstrate the performance of the algorithm on a synthetic example involving electrical resistivity imaging of unimodal and bimodal solute plumes. We show that Bayesian-McMC inversion can proceed in the reduced dimensional model space in order to make it tractable and produce physically realistic solute plumes.

Introduction

Long-term monitoring and characterization of hydrogeological processes are crucial to the efficient management of groundwater resources. While the traditional well-based sampling methods provide valuable insights of the spatial extent of subsurface solute plumes (Freyberg, 1986; LeBlanc et al., 1991), they are invasive, expensive, and pose risks for contaminant mobilization. Consequently, there is growing interest in the use of geophysical techniques to rapidly and non-invasively estimate spatially continuous hydrogeological features.

Traditional hydrogeophysical deterministic inverse methods, however, require regularization constraints for computational stability (Tikhonov and Arsenin, 1977), which typically impose smoothness and/or force the solution toward some reference model (Menke, 1984; Pidlisecky et al., 2007, Johnson et al., 2010) that does not properly represent the physics of the target hydrologic process or properties. As demonstrated by Day-Lewis et al. (2007), the choice of regularization strongly influences the outcome of the inversion procedure. To circumvent the effects of regularization on petrophyiscal transformation, Hermans et al. (2016) applied the prediction-focused approach (Satija and Caers, 2015) for direct prediction without the need for classical inversion of their data.

Unlike the deterministic inversion framework, stochastic imaging (SI) can be solved with or without regularization assumptions, making it more robust and intuitively appealing, particularly, for the characterization of geologically realistic hydrologic properties (e.g., Arpat and Caers, 2004; Oware, 2016). SI recovers multiple realizations that enable uncertainty assessment, which is particularly crucial in view of the limited, noisy measurements coupled with our incomplete understanding of the target hydrogeologic structure. Bayesian Markov-chain-Monte-Carlo (McMC) is a widely used SI strategy in hydrogeophysics (e.g., Irving and Singha, 2010; Cordua et al., 2012). Standard McMC sampling methods, however, can become computationally intractable when working with spatially distributed (high-dimensional) geophysical parameter fields. Recently, multiple researchers (e.g., Ruggeri et al, 2015; Binley et al., 2015; Vrugt, 2016) have noted the potential of performing McMC in a reduced dimensionality space, which offers a viable approach to addressing the intractability of McMC-based inversion strategies when working with high-dimensional problems.

The most widely used statistical tool for dimensionality truncation is the proper orthogonal decomposition (POD) (e.g., Banks et al., 2000; Pinnau, 2008), also called the singular value decomposition (SVD) (e.g, Lawson and Hanson, 1995), principal component analysis (PCA) (Jollife, 2002), or the Karhunen-Loève transform. POD finds an orthogonal set of basis vectors that capture the maximum amount of variability in a training dataset, thereby enabling a sparse representation of the chosen system in a computationally efficient manner, making POD-based inversion strategies model order reduction (MOR) techniques [e.g., Banks et al., 2000; Oware and Moysey, 2014]. While MOR techniques have been extensively researched in other disciplines, such as applied mathematics (e.g., Li et al., 2009; Yao and Meerbergen, 2013) and image processing (e.g., Milanfer et al., 1996), they remain largely under-explored in the field of hydrogeophysics. In contrast to our proposed use of POD as a sparse basis for representing the parameter space, the use of POD in groundwater inverse problems generally tends to focus on improving the computational efficiency of the forward model (e.g., Winton et al., 2011). Some applications of POD/SVD/PCA in the hydrologeology and geophysics literature include Jacobson (1985) and Oware et al. (2012, 2013).

10.1190/segam2018-w12-01.1

Tonkin and Doherty (2005) proposed a hybrid regularized inversion for highly parameterized environmental models based on a combination of SVD and the traditional regularization constraints. Their approach was deterministic and required traditional regularization constraints for computational stability. Furthermore, unlike Oware et al. (2013) who determined the optimal basis vectors from training images, thereby facilitating the incorporation of prior hydrologic process (physics-based) constraints to augment the geophysical measurements, Tonkin and Doherty (2005) estimated the optimal basis vectors from the observational data, in addition to employing the data to condition the inversion. In the stochastic framework, Tompkins et al. (2011) utilized deterministic inversion of their data to estimate the mean and linearized covariance, and then applied PCA on the linearized covariance for posterior sampling around the deterministically estimated mean. The uncertainty quantification in their approach is, therefore, limited to the level of uncertainty evaluated from the deterministically estimated linearized covariance matrix. Laloy et al. (2012b) successfully performed McMC in the lower-dimensional model space related to Legendre moments and predefined mass and morphological features. Their use of predefined mass and morphological constraints in an attempt to conserve mass and produce realistic plume morphologies impose hard constraints that are typically unknown *a priori*. We present a novel Bayesian-McMC strategy based on Oware et al. (2013). The approach uses hydrologic-process-tuned non-parametric basis vectors to constrain the inversion procedure in the lower-dimensional model space.

Basis-Constrained Bayesian-McMC Algorithm

As demonstrated by Oware *et al*. (2013) and Oware and Moysey (2014), an *N*-by-*M* electrical conductivity target distribution ($\boldsymbol{\sigma}$) can be reconstructed as a linear combination of its basis vectors, **B**, and appropriate coefficients, **c**, expressed mathematically as:

$$\boldsymbol{\sigma} = \mathbf{Bc}, \qquad (1)$$

Because the reconstruction in Eq. 1 can be performed with a small number, *p*, of selected basis vectors compared to the size of the full dimensionality space (*N*-by-*M*), the number of inversion parameters p << *N*-by-*M*, making the reconstruction a model-order-reduction technique. Using Eq. 1, Oware et al. (2013) developed the basis-constrained inversion wherein the estimation of the optimal set of coefficients required to weigh the basis to reconstruct the target is conditional on the geophysical measurements. Here, we formulate a basis-constrained Bayesian-McMC inversion framework (a stochastic version of Oware at al., 2013):

$$\mathbf{c}_{post} = \mathbf{c}_{prior} L(\mathbf{d}_{obs}|\boldsymbol{\sigma}) = \mathbf{c}_{prior} L(\mathbf{d}_{obs}|\mathbf{Bc}) \qquad (2)$$

where $\mathbf{c}_{post}$ and $\mathbf{c}_{prior}$ denote the posterior and prior coefficients, respectively. $L(\cdot)$ represents the likelihood function, which assess the fidelity of a proposed model given the observed data, $\mathbf{d}_{obs}$. It should be noted that the basis, **B**, in Eq. 2 is a constant, which makes Eq. 2 a basis-constrained Bayesian-McMC inversion strategy. A projection of the posterior coefficients into the optimal basis space will produce the posterior conductivity realizations, $\boldsymbol{\sigma}_{post}$:

$$\boldsymbol{\sigma}_{post} = \mathbf{Bc}_{post} \qquad (3)$$

The workflow of the basis-constrained Bayesian McMC proceeds as follows (Fig. 1): First, we perform Monte Carlo simulations to generate training images (TIs) tuned to the target hydrologic process. Second, we construct basis vectors, **B**, from the TIs. Third, to obtain prior distributions of the coefficients, $\mathbf{c}_{prior}$, we project the TIs onto **B** to obtain coefficients structure associated with each TI. The estimated mean and standard deviations of the prior coefficients are employed to define prior Gaussian distributions. Fourth, to propose coefficients for the McMC procedure, we re-simulate the coefficients from the prior

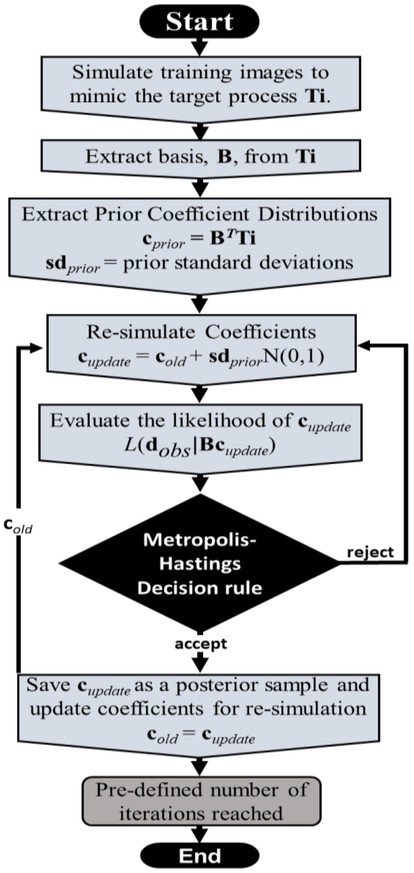
Start
Simulate training images to mimic the target process **Ti**.
Extract basis, **B**, from **Ti**
Extract Prior Coefficient Distributions
$\mathbf{c}_{prior} = \mathbf{B}^T\mathbf{Ti}$
$\mathbf{sd}_{prior}$ = prior standard deviations
Re-simulate Coefficients
$\mathbf{c}_{update} = \mathbf{c}_{old} + \mathbf{sd}_{prior}\mathrm{N}(0,1)$
Evaluate the likelihood of $\mathbf{c}_{update}$
$L(\mathbf{d}_{obs}|\mathbf{Bc}_{update})$
$\mathbf{c}_{old}$
Metropolis-Hastings Decision rule
reject
accept
Save $\mathbf{c}_{update}$ as a posterior sample and update coefficients for re-simulation
$\mathbf{c}_{old} = \mathbf{c}_{update}$
Pre-defined number of iterations reached
End

Figure 1: Schematic of the basis-constrained Bayesian McMC algorithm.

Gaussian distributions. We accept or reject the proposed coefficients based on the classical Metropolis-Hastings acceptance rule (Metropolis et al., 1953; Hastings, 1970). Finally, the optimal coefficients are projected onto the basis (Eq. 3) to obtain multiple realizations of the target.

Numerical Demonstration of the Algorithm

We demonstrate the performance of the algorithm with geoelectrical imaging of two synthetic reference models, unimodal and bimodal solute plumes. We use the same reference models by Oware et al. (2013). However, while the morphologies of the plumes are maintained, the simulation domain is re-scaled to a 100 cm by 30 cm to match the dimensions of a lab-scale sandbox (lab-scale results not shown here). The unimodal and bimodal plumes represent single and double source flow and solute transport scenarios, respectively.

We also use the same 400 training images (TIs) by Oware at al. (2013). The generation of the TIs assumed a single source hydrologic transport process consistent with the unimodal test case. While bimodality in plume morphology was not conceptualized, we employ the same set of TIs to reconstruct the bimodal target. This was done to test the flexibility of the algorithm to reconstruct a target at a site with limited prior knowledge about the nature of the prevailing transport process. We construct the basis vectors from the TIs using proper orthogonal decomposition (POD). Fig. 2 shows the first 20 principal basis vectors obtained from the 400 TIs that are used to constrain all the inversion results presented here.

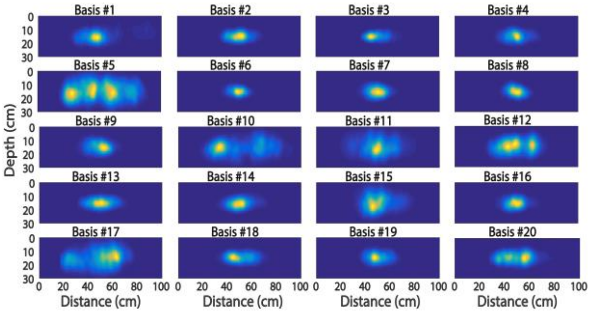

Figure 2: First 20 principal POD basis (POD plumes) constructed from training images under the assumption of a single source transport process.

For the electrical resistivity experiment, we deployed 32 surface electrodes with 3 cm electrode spacing. We used circulating dipole-dipole to acquire a total of 311 quadrupoles at the end of each survey. To mimic real-world data, the observed data were corrupted with an uncorrelated Gaussian noise with standard deviation proportional to 3% of the data values. We used the MATLAB based resistivity forward simulation code FW2_5D (Pidlisecky and Knight, 2008) for all resistivity forward simulations.

During the reconstruction, the range of the sampling coefficients is crucial to obtaining physically realistic solute plumes. As posited by Tompkins et al. (2011), while every model has a unique mapping in the optimal reduced space via the coefficients (Eq. 1), not every set of coefficients will produce feasible models. Hence, to produce physically realistic models, the sampling of the coefficients must be bounded. To accomplish this, Tompkins et al. (2011) defined individual bounds for the coefficients and employed constraint mapping (Ganapathysubramanian and Zabaras, 2007) with optimization to estimate the feasible range of each coefficient. Their approach lacks information about the physics of the target process that defines the feasible model space. Here, to produce physics-based prior distributions for the coefficients, we map the TIs into the reduced **B** space, i.e.:

$$\mathbf{c}_{prior} = \mathbf{B}^T\mathbf{T}_i, \qquad (4)$$

where *T* represents transpose and $\mathbf{T}_i$ is the set of training images. From the mapping in Eq. 4, there is a unique set of coefficients associated with each TI. For the 400 TIs considered here, for instance, there are 400 realizations of each coefficient. Fig. 3 displays the histograms of the coefficients constructed from the 400 realizations for the first 20 principal POD plumes shown in Fig. 2. The mean and standard deviation of each coefficient defines a prior Gaussian distribution for resampling the coefficients. The fairly Gaussian nature of the histograms (Fig. 3) justify our assumption of prior Gaussian distributions for the coefficients. For a severely non-Gaussian behavior, however, the bootstrapping resampling approach (Mooney et al., 1993) can be employed for non-parametric resimulation of the coefficients.

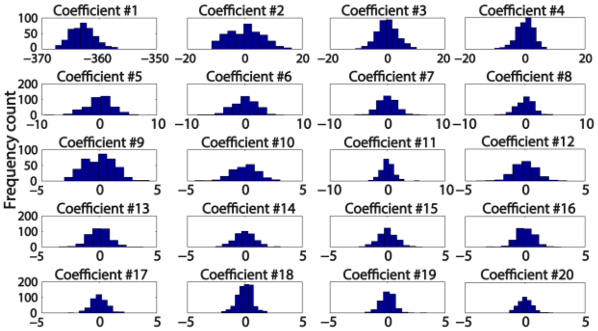

Figure 3: Histograms of prior coefficients ($\mathbf{c}_{prior}$) corresponding to the first 20 principal POD plumes shown in Fig. 3.

Results and Discussion

We ran the algorithm for 300,000 iterations for both the unimodal and bimodal test cases. In all the reconstructions presented here, we estimated 300 coefficients compared to the 3,000 model parameters, resulting in 90% truncation in the dimensionality of the problem.

10.1190/segam2018-w12-01.1

Examination of the sampling paths of the data misfits (not shown) reveals that the algorithm burned-in at about 30,000 iterations, resulting in 270,000 posterior samples. To acquire uncorrelated samples to prevent uncertainty underestimation in the posterior analysis, we performed autocorrelation analysis (e.g., Dorman et al., 2007) to find the number of iterations required to generate independent samples. To estimate a more representative autocorrelation length, we performed the autocorrelation analysis for multiple samples and estimated the average. Analysis of the evolution of the average correlation coefficient (Fig. 4) indicates that the algorithm generates independent samples after about every 3,000 iterations, resutling in 90 uncorrelated samples for the posterior analysis. While Fig. 4 is based on the unimodal scenario, similar results were obtained in the bimodal test case.

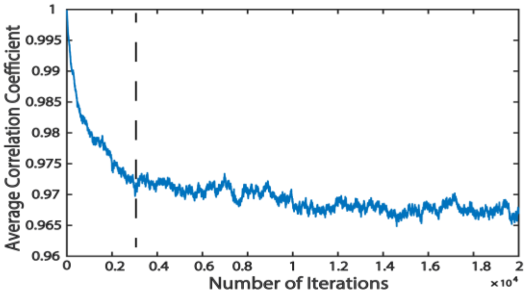

Figure 4: Evolution of the average correlation coefficients. The dashed line marks the iteration at which linear correlation was lost between consecutive posterior samples, representing the indepdent sampling correlation length.

The inversion results for the unimodal test case are presented in Fig. 5. Qualitative comparison of the target plume (Fig. 5A) with the posterior mean estimate (Fig. 5B) shows that the target plume was highly and compactly resolved with minimal smearing. Visual inspection of the three posterior samples (Figs. 5D-F) reveals that the posterior samples are idependent and identical. Inspection of the standard deviation panel (Figs. 5C) unravels high uncertainty associated with the estimation of the high conductivity region along the mid-section of the plume. The pattern in the uncertainty estimation is consistent with the fact that the peak conductivity region along the mid-section of the plume appears underestimated, which highlights the utility of uncertainty quantification in the interpretation of geophysical tomograms.

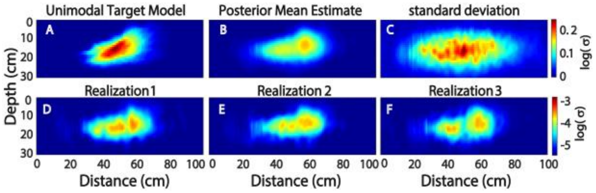

Figure 5: Comparison of the true unimodal plume (A) with the posterior mean estimate (B) and three idependent and identical realizations (D-F). Note that the standard deviation panel (C) is on a different color scale.

Fig. 6 outlines the inversion results for the bimodal test case. Visual comparison of the bimodal plume (Fig. 6A) with the posterior mean estimate (Fig. 6B) shows that while the top plume seems highly constrained, the size of the bottom plume appears underestimated. The underestimation of the bottom plume is attributable to the poor data sensitivity with increasing depth for surface electrical resistivity measurements. Once again, a close look at the three posterior samples (Figs. 6D-F) shows that the posterior realizations are independent and identical. Examination of the standard deviation panel (Fig. 6C) points to high uncertainty related to the estimation of the low conductivity region between the two plumes. Qualitative comparsion of the target and the posterior mean estimate shows that the low conductivity region between the two plumes appears to be the most poorly resolved area in the estimation, which corroborates the uncetainty pattern depicted in the uncertainty estimation. Nevertheless, although bimodality was not conceptualized, the algorithm was till able to reconstruct the bimodality in the target.

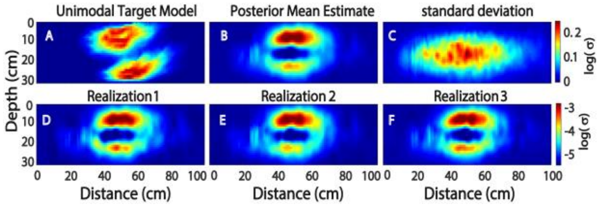

Figure 6: Comparison of the true bimodal plume (A) with the posterior mean estimate (B) and three idependent and identical realizations (D-F). Note that the scenario used to produce the bimodal plume had two distinct source zones for the conductive solute, whereas the conceptual model used to produce the inversion results all assumed a single source zone.

Conclusion

Geophysical imaging is becoming increasingly indispensable to non-invasively gain insight into subsurface processes. This is critical to monitoring transport of contaminants, designing and evaluating the performance of *in situ* remediation schemes, and to facilitate decision making by regulatory bodies. While uncertainty quantification is crucial for comprehensive interpretation of geophysical estimations, the standard McMC method can become computationally prohibitive and intractable, particularly when it comes to the estimation of physically realistic solute plumes. Proposed here is a novel Bayesian-McMC inversion strategy that proceeds in the reduced dimensionality space, thereby reducing the number of inversion parameters to be estimated (perturbed). Our approach uses hydrologic-process tuned non-parametric basis vectors to constrain the inversion procedure, resulting in the estimation of physically realistic solute plumes.

10.1190/segam2018-w12-01.1

REFERENCES

Arpat, G. B., and J. Caers, 2004, Multiple-scale, pattern-based approach to sequential simulation, *in* O. Leuangthong, and C. V. Deutsch, eds., Geostatistics Banff 2004, 255–254.

Banks, H. T., M. L. Joyner, B. Wincheski, and W. P. Winfree, 2000, Nondestructive evaluation using a reduced-order computational methodology: Inverse Problems, **16**, 929–945, https://doi.org/10.1088/0266-5611/16/4/304.

Binley, A., S. S. Hubbard, J. A. Huisman, A. Revil, D. A. Robinson, K. Singha, and L. D. Slater, 2015, The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales: Water Resources Research, **51**, 3837–3866, https://doi.org/10.1002/2015wr017016.

Cordua, K. S., T. M. Hansen, and K. Mosegaard, 2012, Monte Carlo full-waveform inversion of crosshole GPR data using multiple-point geostatistical a priori information: Geophysics, **77**, no. 2, H19–H31, https://doi.org/10.1190/geo2011-0170.1.

Day-Lewis, F. D., Y. Chen, and K. Singha, 2007, Moment inference from tomograms: Geophysical Research Letters, **34**, https://doi.org/10.1029/2007GL031621.

Freyberg, D. L., 1986, A natural gradient experiment on solute transport in a sand aquifer: 2. Spatial moments and the advection and dispersion of nonreactive tracers: Water Resources Research, **22**, 2031–2046, https://doi.org/10.1029/WR022i013p02031.

Ganapathysubramanian, B., and N. Zabaras, 2007, Modeling diffusion in random heterogeneous media: Data-driven models, stochastic collocation and the variational multiscale method: Journal of Computational Physics, **226**, 326–353, https://doi.org/10.1016/j.jcp.2007.04.009.

Hastings, W., 1970, Monte Carlo sampling methods using Markov chains and their applications: Biometrika, **57**, 97, https://doi.org/10.2307/2334940.

Hermans, T., E. K. Oware, and J. K. Caers, 2016, Direct prediction of spatially and temporally varying physical properties from time-lapse electrical resistance data: Water Resources Research, **52**, 7262–7283, https://doi.org/10.1002/2016WR019126.

Irving, J., and K. Singha, 2010, Stochastic inversion of tracer test and electrical geophysical data to estimate hydraulic conductivities: Water Resources Research, **46**, https://doi.org/10.1029/2009wr008340.

Jacobson, E. A., 1985, A statistical parameter estimation method using singular value decomposition with application to Avra Valley aquifer in southern Arizona.

Johnson, T. C., R. J. Versteeg, A. Ward, F. D. Day-Lewis, and A. Revil, 2010, Improved hydrogeophysical characterization and monitoring through parallel modeling and inversion of time-domain resistivity and induced polarization data: Geophysics, **75**, no. 4, 27–41, https://doi.org/10.1190/1.3475513.

Jolliffe, I., 2002, Principal component analysis: John Wiley & Sons, Ltd.

Laloy, E., N. Linde, and J. A. Vrugt, 2012, Mass conservative three-dimensional water tracer distribution from Markov chain Monte Carlo inversion of time-lapse ground penetrating radar data: Water Resources Research, **48**, https://doi.org/10.1029/2011wr011238.

Lawson, C. L., and R. J. Hanson, 1974, Solving least squares problems: Prentice-Hall.

Leblanc, D. R., 1991, Large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachusetts. 1. Experimental design and observed tracer movement: NTIS.

Li, Y.-T., Z. Bai, and Y. Su, 2009, A two-directional arnoldi process and its application to parametric model order reduction: Journal of Computational and Applied Mathematics, **226**, 10–21, https://doi.org/10.1016/j.cam.2008.05.059.

Menke, W., 1984, Geophysical data analysis: Discrete Inverse Theory: Academic Press, London, p. 289.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, 1953, Equation of state calculations by fast computing machines: The Journal of Chemical Physics, **21**, 1087–1092, https://doi.org/10.1063/1.1699114.

Milanfar, P., W. C. Karl, and A. S. Willsky, 1996, A moment-based variational approach to tomographic reconstruction: IEEE Transactions on Image Processing, **5**, 459–470.

Oware, E. K., 2016, Estimation of hydraulic conductivities using higher-order MRF-based stochastic joint inversion of hydrogeophysical measurements: The Leading Edge, **35**, 776–785, https://doi.org/10.1190/tle35090776.1.

Oware, E. K., and S. M. J. Moysey, 2014, Geophysical evaluation of solute plume spatial moments using an adaptive POD algorithm for electrical resistivity imaging: Journal of Hydrology, **517**, 471–480, https://doi.org/10.1016/j.jhydrol.2014.05.054.

Oware, E., S. Moysey, and T. Khan, 2013, Physically based regularization of hydrogeophysical inverse problems for improved imaging of process-driven systems: Water Resources Research, **49**, 6238–6247, https://doi.org/10.1002/wrcr.20462.

Oware, E. K., S. M. J. Moysey, and T. Khan, 2012, Improved imaging of electrically conductive solute plumes using a new strategy for physics based regularization of resistivity imaging problems: 82nd Annual International Meeting, SEG, Expanded Abstracts, https://doi.org/10.1190/segam2012-1367.1.

Pidlisecky, A., E. Haber, and R. Knight, 2007, A 3D resistivity inversion package: Geophysics, **72**, no. 2, H1–H10, https://doi.org/10.1190/1.2402499.

Pidlisecky, A., and R. J. Knight, 2008, FW2_5D: A MATLAB 2.5-D electrical resistivity modeling code: Computers & Geosciences, **34**, 1645–1654, https://doi.org/10.1016/j.cageo.2008.04.001.

Pinnau, R., 2008, Model reduction via proper orthogonal decomposition. Model Order Reduction: Theory, Research Aspects and Applications: Springer, 95–109.

Ruggeri, P., J. Irving, and K. Holliger, 2015, Systematic evaluation of sequential geostatistical resampling within MCMC for posterior sampling of near-surface geophysical inverse problems: Geophysical Journal International, **202**, 961–975, https://doi.org/10.1093/gji/ggv196.

Satija, A., and J. Caers, 2015, Direct forecasting of subsurface flow response from non-linear dynamic data by linear least-squares in canonical functional principal component space: Advances in Water Resources, **77**, 69–81, https://doi.org/10.1016/j.advwatres.2015.01.002.

Tikhonov, A. N., and V. Y. Arsenin, 1977, Solutions of ill-posed problems: John Wiley & Sons.

Tompkins, M. J., J. L. Fernández Martínez, D. L. Alumbaugh, and T. Mukerji, 2011, Scalable uncertainty estimation for nonlinear inverse problems using parameter reduction, constraint mapping, and geometric sampling: Marine controlled source electromagnetic examples: Geophysics, **76**, no. 4, F263–F281, https://doi.org/10.1190/1.3581355.

Tonkin, M. J., and J. Doherty, 2005, A hybrid regularized inversion methodology for highly parameterized environmental models: Water Resources Research, **41**, https://doi.org/10.1029/2005wr003995.

Vrugt, J. A., 2016, Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation: Environmental Modelling & Software, **75**, 273–316, https://doi.org/10.1016/j.envsoft.2015.08.013.

Yao, Y., and K. Meerbergen, 2013, Parametric model order reduction of damped mechanical systems via the block Arnoldi process: Applied Mathematics Letters, **26**, 643–648, https://doi.org/10.1016/j.aml.2013.01.006.

10.1190/segam2018-w12-01.1