Dimensionality reduction for efficient Bayesian estimation of groundwater flow in strongly heterogeneous aquifers

Thierry A. Mara, Noura Fajraoui, Alberto Guadagnini, Anis Younes

Dimensionality reduction for efficient Bayesian estimation of groundwater flow in strongly heterogeneous aquifers

Dimensionality reduction for efficient Bayesian estimation of groundwater flow in strongly heterogeneous aquifers

Thierry A. Mara[a], Noura Fajraoui[b,c], Alberto Guadagnini[d,e], Anis Younes[b,f,g*]

Full reference: Mara, T.A., N. Fajraoui, A. Guadagnini, and A. Younes (2017), Dimensionality reduction for efficient Bayesian estimation of groundwater flow in strongly heterogeneous aquifers, *Stoch. Environ. Res. Risk Assess.*, 31, 2313-2326, doi:10.1007/s00477-016-1344-1.

[a] PIMENT, EA 4518, Université de La Réunion, FST, 15 Avenue René Cassin, 97715 Saint-Denis, Réunion

[b] LHyGeS, UMR-CNRS 7517, Université de Strasbourg/EOST, 1 rue Blessig, 67084 Strasbourg, France

[c] Chair of Risk, Safety and Uncertainty Quantification, Department of Civil Engineering, ETH Zurich, Switzerland

[d] Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Milano, Italy.

[e] Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, Arizona, USA

[f] IRD UMR LISAH, F-92761 Montpellier, France

[g] LMHE, Ecole Nationale d'Ingénieurs de Tunis, Tunisie

Submitted to *Stochastic Environmental Research and Risk Assessment*

* Contact person: Anis Younes

E-mail: younes@unistra.fr

Abstract

We focus on the Bayesian estimation of strongly heterogeneous transmissivity fields conditional on data sampled at a set of locations in an aquifer. Log-transmissivity, Y, is modeled as a stochastic Gaussian process, parameterized through a truncated Karhunen-Loève (KL) expansion. We consider Y fields characterized by a short correlation scale as compared to the size of the observed domain. These systems are associated with a KL decomposition which still requires a high number of parameters, thus hampering the efficiency of the Bayesian estimation of the underlying stochastic field. The distinctive aim of this work is to present an efficient approach for the stochastic inverse modeling of fully saturated groundwater flow in these types of strongly heterogeneous domains. The methodology is grounded on the construction of an optimal sparse KL decomposition which is achieved by retaining only a limited set of modes in the expansion. Mode selection is driven by model selection criteria and is conditional on available data of hydraulic heads and (optionally) Y. Bayesian inversion of the optimal sparse KLE is then inferred using Markov Chain Monte Carlo (MCMC) samplers. As a test bed, we illustrate our approach by way of a suite of computational examples where noisy head and Y values are sampled from a given randomly generated system. Our findings suggest that the proposed methodology yields a globally satisfactory inversion of the stochastic head and Y fields. Comparison of reference values against the corresponding MCMC predictive distributions suggests that observed values are well reproduced in a probabilistic sense. In a few cases, reference values at some unsampled locations (typically far from measurements) are not captured by the posterior probability distributions. In these cases, the quality of the estimation could be improved, e.g., by increasing the number of measurements and/or the threshold for the selection of KL modes.

Keywords: Heterogeneous porous media; Stochastic inverse modeling; Karhunen-Loève expansion; Markov Chain Monte Carlo

1. Introduction

Prediction of flow and transport in subsurface reservoirs is typically fraught with diverse types of uncertainties, including imperfect knowledge of the spatial distribution of system parameters, types of boundary conditions and their values, as well as forcing terms (e.g. Lin et al. 2010; Tartakovsky et al. 2012; Tartakovsky 2013 and references therein). All these uncertainties should be appropriately considered and their impact on the quality of model predictions needs to be quantified in a rigorous way. These requirements should also be compatible with the operational challenges associated with the analysis and management of complex settings such as those characterizing natural aquifer systems.

Bayesian inference is a convenient and flexible theoretical framework within which all these issues can be tackled. Bayesian approaches enable one to incorporate in a stochastic model inversion available data from diverse sources, relying on prior information. The latter is then updated through conditioning onto observations to yield posterior probability distributions of system parameters and responses. Recent examples involving applications of Bayesian characterizations of uncertain parameter fields associated with subsurface flow and transport settings can be found, among others, in Rubin et al. (2010), Murakami et al. (2010), Chen et al. (2012), and Over et al. (2013) and references therein.

The application of the Bayesian framework to (stochastic) inverse modeling of groundwater flow typically requires obtaining multiple forward solutions of the mathematical model governing the spatial/temporal evolution of the system physics. The Markov Chain Monte Carlo (MCMC) method is one of the most widely employed approaches in the context of porous media characterization. MCMC has been applied with several degrees of success in hydrogeology for stochastic model calibration and uncertainty quantification (e.g., Vrugt et al. 2003, 2008; Zanini and Kitanidis 2009; Keating et al. 2010; Schoups and Vrugt 2010; Huard et al. 2010; Zheng and Han 2016). Shi et al. (2012) employed MCMC for vadose zone

characterization and compared the ensuing results against those obtained through a nonlinear regression method. These authors found that MCMC (*a*) produces results of higher fidelity and (*b*) is more advantageous from a computational standpoint than nonlinear regression for problems associated with a relatively small dimensionality of the parameter space.

Routine application of MCMC to stochastic inverse groundwater flow modeling under realistic conditions is hampered by practical challenges due to the usually high dimensionality of the parameter space. Parameterization of the spatially heterogeneous distribution of model attributes, such as system transmissivity, via the truncated Karhunen-Loève Expansion (KLE) (Loève 1977) can be considered as a viable strategy to alleviate this difficulty. In essence, the Karhuen-Loève representation of a random spatial field is based on the spectral expansion of the process covariance function. This approach has been broadly used (Li and Cirpka 2006; Efendiev et al. 2006; Marzouk and Najm 2009; Ray et al. 2012; Laloy et al. 2013; Mara et al. 2015) mainly because it enables one to reduce the dimensionality of the problem while preserving to a given extent the key characteristics of the considered stochastic model (Marzouk and Najm 2009). The KLE has been recently used by Das et al. (2010) in conjunction with the MCMC technique to characterize the saturated hydraulic conductivity of a mildly heterogeneous agricultural field. These authors rely on a truncated form of KLE by retaining solely a reduced number of terms (or modes) in the expansion.

The number of terms that enables the truncated KLE to be effective for a computationally affordable and accurate system representation depends on the functional format of the covariance function (e.g., exponential, Gaussian, spherical, or other) as well as on the degree of spatial persistence, or correlation, of the field. It can be seen that the norm of the eigenvalues of the covariance matrix tends to decay rapidly for heterogeneous fields characterized by large correlation scales (relative to a characteristic length scale of the flow domain). In these cases, it is seen that retaining less than 20 terms in the KLE typically allows

capturing more than 90% of the energy of the target spatial random field (Das et al. 2010). Otherwise, the number of terms to be retained in the KLE to achieve an appropriate representation of a random parameter field tends to increase when the correlation scale of the covariance function decreases. This can become a limiting factor constraining the effectiveness of the technique when one is confronted with short-range (with respect to the domain size) correlated heterogeneous fields.

In this work, we focus on these types of strongly heterogeneous fields, for which Bayesian inference becomes highly challenging and computationally demanding due to the large number of terms required to be retained in the KLE. The main objective of this work is to develop an operational strategy which renders the MCMC method computationally affordable to be employed for the stochastic characterization of short-range random parameter fields. Our strategy is data-driven and is based on destructuring the stochastic inverse modeling procedure of fully saturated groundwater flow into the following two steps:

1. Starting from a highly-parameterized system, a set of sparse KLEs are formed by progressively reducing the dimensionality of the parameter space. For each KLE, the MAximum a Posteriori (MAP) estimate of the eigenmodes in the expansion is obtained through inverse modeling of flow (against available observations of the system state, i.e., hydraulic heads or fluxes, and, optionally, of system parameters, i.e., hydraulic conductivity/transmissivity). Once this MAP estimate is obtained, a new sparse KLE is constructed by removing the least influential components of the expansion via an analysis of the spatial variance of the resulting estimated field.
2. A model selection criterion is employed to select the optimal sparse KLE, as driven by the available data. The posterior statistical distribution of the corresponding eigenmodes is then obtained, relying on the DREAM$_{(ZS)}$ MCMC sampler developed by Laloy and Vrugt (2012).

The work is organized as follows: Section 2 introduces the flow problem and Section 3 the Karhunen-Loève decomposition. In Section 4, we detail the way the Bayesian inference is performed for a stochastic field of the kind we consider in our computational example. Section 5 summarizes the main elements of the information criterion we employ for model selection. Section 6 illustrates our strategy to achieve dimensionality reduction of the parameter space. Section 7 is devoted to the presentation of an application of our technique to the stochastic inversion of flow through a strongly heterogeneous random porous medium. The key findings are then summarized in the conclusions.

2. The flow model

We consider two-dimensional steady-state fully saturated groundwater flow taking place within a spatially bounded domain, *D*, governed by

$$\begin{cases} \nabla.\left(T\left(\boldsymbol{x}\right)\nabla h\left(\boldsymbol{x}\right)\right)=0, & \boldsymbol{x}\in D \\ h\left(\boldsymbol{x}\right)=h_0, & \boldsymbol{x}\in \partial D_1 \\ \left(-T\left(\boldsymbol{x}\right)\nabla h\left(\boldsymbol{x}\right)\right).\boldsymbol{\eta}_{\partial D_2}=g_0 & \boldsymbol{x}\in \partial D_2 \end{cases} \tag{1}$$

Here, $\boldsymbol{x}=\left(x,y\right)$ is the vector of spatial coordinates, $h\left(\boldsymbol{x}\right)$ [L] and $T\left(\boldsymbol{x}\right)$ [L^2T^{-1}] respectively are hydraulic head and transmissivity fields; Dirichlet and Neumann boundary conditions corresponding to given pressure head, h_0, or normal flux, g_0, are respectively defined along the (disjoint) boundary segments ∂D_1 and ∂D_2, forming the domain boundary ∂D; $\boldsymbol{\eta}_{\partial D_2}$ is the outward unit vector normal to ∂D_2.

Given the spatial distribution of $T\left(\boldsymbol{x}\right)$, the numerical solution of the forward problem (1) is performed through the mixed-hybrid finite element method (Younes et al. 2010) upon discretizing *D* with uniform square elements.

Observations of $h\left(\boldsymbol{x}\right)$ and $T\left(\boldsymbol{x}\right)$ are assumed to be jointly available at a set of *M* points $\boldsymbol{x}_i=\left(x_i,y_i\right)$ (*i* = 1, 2, ..., *M*) within *D*. We collect these data into the observation vector $\boldsymbol{m}$.

For the purpose of our demonstration we assume that the functional format of the covariance of $Y(\boldsymbol{x}) = log(T(\boldsymbol{x}))$ is deterministically known together with its parameters. We consider log-transmissivity Y as a Gaussian field that can be represented by its Karhunen-Loève expansion (Loève 1977).

3. Karhunen-Loève expansion

Let $Y(\boldsymbol{x},\omega) = log(T(\boldsymbol{x},\omega))$ be a Gaussian random process, where $\boldsymbol{x} \in D$ and $\omega \in \Omega$ (Ω being a suitable probability space). One can characterize Y through its mean, μ_Y, and two-point covariance function, $C_Y(\boldsymbol{x}, \boldsymbol{x}')$, between locations $\boldsymbol{x}$ and $\boldsymbol{x}'$. Covariance C_Y is bounded, symmetric, and positive definite (assuming that $Y \in L^2(D), \forall \boldsymbol{x} \in D$). The Karhunen-Loève Expansion (KLE) of the random field $Y(\boldsymbol{x},\omega)$ is defined as

$$Y(\boldsymbol{x},\omega) \equiv \mu_Y + \sum_{i=1}^{+\infty} \sqrt{\lambda_i}\,\xi_i(\omega)\,\varphi_i(\boldsymbol{x}) \tag{2}$$

Here, λ_i and $\varphi_i(\boldsymbol{x})$ respectively are eigenvalues and eigenfunctions of $C_Y(\boldsymbol{x}, \boldsymbol{x}')$, $\{\xi_i\}_{i=1}^{\infty}$ being a set of statistically independent standard normal random variables. According to Mercer's theorem (Mercer, 1909) $C_Y(\boldsymbol{x}, \boldsymbol{x}')$ can be decomposed as

$$C_Y(\boldsymbol{x},\boldsymbol{x}') = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\boldsymbol{x})\,\varphi_i(\boldsymbol{x}') \tag{3}$$

where λ_i and $\varphi_i(\boldsymbol{x})$ are obtained by solving the following Fredholm equation

$$\int_D C_Y(\boldsymbol{x},\boldsymbol{x}')\,\varphi_i(\boldsymbol{x}')\,d\boldsymbol{x}' = \lambda_i \varphi_i(\boldsymbol{x}). \tag{4}$$

The eigenfunctions $\{\varphi_i(\boldsymbol{x})\}_{i=1}^{\infty}$ are orthonormal and form a complete basis in $L^2(D)$, i.e.,

$$\int_D \varphi_i(\boldsymbol{x})\,\varphi_j(\boldsymbol{x})\,d\boldsymbol{x} = \delta_{ij} \tag{5}$$

δ_{ij} being the Kronecker delta.

The separability assumption is often used to characterize the covariance function model of Y in the context of stochastic analyses of flow and transport in randomly heterogeneous porous and/or fractured formations. This assumption has enabled obtaining analytical solutions of key moments of hydraulic head and fluxes and contaminat transport and facilitates basic studies of uncertainty propagation in such random porous and fractured media (see. e.g. Dagan 1989; Zhang 2002, and references therein). Adoption of this simplified format has also the practical advantage of being associated with relatively straightforward estimates of the model parameters through the type and quantity of data which is typically available (see e.g. Gneiting et al. 2007, Genton 2007). In the following, we assume that the covariance function of $Y(\boldsymbol{x},\omega)$ has the exponential form

$$C_Y(\boldsymbol{x},\boldsymbol{x}') = \sigma^2 \exp\left(-\frac{|x-x'|}{\eta} - \frac{|y-y'|}{\eta}\right) \tag{6}$$

where σ^2 and η respectively are the variance and correlation length of *Y*. The eigenvalues λ_i and corresponding eigenfunctions appearing in (2)-(5) can be readily computed (Zhang and Lu 2004) by solving a system of two coupled algebraic equations. In the most general case, the eigenvalue problem (4) is solved numerically (e.g., Phoon et al. 2002). Note that other models could be employed for the representation of C_Y, including, e.g., the Modified Exponential and the Spartan covariance (e.g. Spanos et al. 2007, Tsantili and Hristopulos 2016, Su and Lucor 2006), which might require a smaller number of KL terms than the exponential covariance (Spanos et al. 2007).

As shown in Zhang and Lu (2004), values λ_i monotonically decrease at the rate of $1/i^2$. One can then approximate $Y(\boldsymbol{x},\omega)$ by considering a finite number of terms in (2), i.e.,

$$Y(\boldsymbol{x},\omega) \approx \mu_Y + \sum_{i=1}^{K} \sqrt{\lambda_i}\,\xi_i(\omega)\,\varphi_i(\boldsymbol{x}) \tag{7}$$

with $\boldsymbol{\xi} \sim \boldsymbol{N}(0, \boldsymbol{I}_K)$, $\boldsymbol{I}_K$ being the identity matrix of size *K*. We note that

$$\sum_{i=1}^{+\infty} \lambda_i = \bar{D}\sigma^2 \tag{8}$$

$\bar{D}$ being a measure of the area of the domain. Hence, the number of terms to be retained in (7) can be selected in a way that the ratio

$$e(K) = \frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{\infty} \lambda_i} \tag{9}$$

is larger than a given threshold. In our computational examples we follow Das et al. (2010) and set $e(K) > 0.90$, which allows to capture more than 90% of the variance of *Y*.

The number of terms to be retained in (7) depends on the correlation length of the covariance function of *Y*, small values of η usually corresponding to high values of *K*. As such, strongly heterogeneous stochastic fields, which are associated with high variance and/or small correlation lengths, pose a clear challenge for an effective representation grounded on the KLE.

The forward problem is tackled by solving (1) for several realizations of the *Y* spatial field. These are obtained by evaluating (7) through sampling of the random vector $\{\xi_i\}_{i=1}^{K}$ from the standard multi-Gaussian distribution. An uncertainty analysis of the way the randomness of *Y* propagates to the output of the flow model can then be easily performed through numerical Monte Carlo simulations. In the context of a stochastic inverse problem, one is mainly interested in characterizing a collection of *Y* fields that are consistent with the observations grouped in vector $\boldsymbol{m}$. When the stochastic inverse problem is set in a Bayesian framework, the posterior (updated) probability density function (pdf) of the field $Y(\boldsymbol{x}, \omega)$ is typically inferred on the basis of available data and prior knowledge about the system.

4. Bayesian inference and Markov Chain Monte Carlo (MCMC) sampling

Characterizing the posterior pdf of $Y(\boldsymbol{x},\omega)$ in the context of Bayesian inference is tantamount to assessing the joint posterior pdf of the entries of the random vector $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^{K}$. The conditional posterior distribution of $\boldsymbol{\xi}$ is defined as

$$p(\boldsymbol{\xi}|\boldsymbol{m}) \propto p(\boldsymbol{m}|\boldsymbol{\xi})\, p(\boldsymbol{\xi}) \tag{10}$$

Here, $p(\boldsymbol{m}|\boldsymbol{\xi})$ is the likelihood function and $p(\boldsymbol{\xi})$ is the prior probability density function of $\boldsymbol{\xi}$, which encapsulates any prior knowledge about the log-transmissivity field. As stated in Section 3, we consider $Y(\boldsymbol{x},\omega)$ as a Gaussian process with the covariance function defined in (6). It then follows that $p(\boldsymbol{\xi}) \sim N(0, \boldsymbol{I}_K)$.

The conditional posterior distribution (10) can be characterized through diverse numerical methods. Markov Chain Monte Carlo (MCMC) samplers are particularly suited for this task. There are several MCMC algorithms proposed in the literature (e.g., Haario et al. 2001; Green and Mira 2001; ter Braak and Vrugt 2008; Vrugt et al. 2009a; Laloy and Vrugt 2012), all of which relying on the Metropolis-Hasting algorithm. In the latter, a new candidate value for parameter $\boldsymbol{\xi}^i$ is generated at the *i*th iteration from a proposal distribution $q(\boldsymbol{\xi}^i|\boldsymbol{\xi}^{i-1})$. Acceptance or rejection of a new candidate is based on the associated Hasting ratio, defined as

$$\alpha = \min\left(1, \frac{p(\boldsymbol{\xi}^i|\boldsymbol{m})\, q(\boldsymbol{\xi}^i|\boldsymbol{\xi}^{i-1})}{p(\boldsymbol{\xi}^{i-1}|\boldsymbol{m})\, q(\boldsymbol{\xi}^{i-1}|\boldsymbol{\xi}^i)}\right) \tag{11}$$

Convergence of the chain to the target distribution, i.e., $p(\boldsymbol{\xi}|\boldsymbol{m})$, is typically achieved after a burn-in period. Considerable research efforts on improving the efficiency of MCMC samplers have been focused on reducing the burn-in period (see, e.g. Haario et al. 2001; Green and Mira 2001; Vrugt et al. 2009a among others). The choice of the proposal distribution $q(.|.)$

and the updating strategy are key to obtain the speed up of the algorithm convergence. A common strategy which is also pursued to accelerate convergence of the MCMC sampler relies on characterizing the modes of the posterior pdf $p\left(\boldsymbol{\xi}\left|\boldsymbol{m}\right.\right)$ (Vrugt and Bouten 2002). Assuming a unimodal pdf, the mode corresponds to the MAximum A Posteriori (MAP) value, defined as

$$\boldsymbol{\xi}^{MAP} = \arg\max_{\xi}\left(p\left(\boldsymbol{\xi}\left|\boldsymbol{m}\right.\right)\right) \tag{12}$$

The MAP characterization enables the MCMC sampler to be initialized approximately around the most likely values associated with the posterior distribution of the model parameter set (Vrugt and Bouten 2002).

Here, we employ the DREAM$_{(ZS)}$ software to generate samples from the conditional posterior distribution of $\boldsymbol{\xi}$ (Laloy and Vrugt 2012). This adaptive algorithm runs multiple chains in parallel to explore the random parameter space. Vrugt et al. (2009b) compared the DREAM algorithm with the generalized likelihood uncertainty estimation (GLUE) method. As a key feature, DREAM$_{(ZS)}$ generates candidates by sampling from an archive of past states collected in a sample **Z**. Thus, only a few parallel chains are required for posterior sampling and a marked reduction of the burn-in period is achieved. The efficiency of the algorithm has been successfully tested on several highly dimensional, complex and nonlinear problems. These studies pointed out that the computational effort can be demanding in cases where the process model be associated with long simulation times. In these instances one can consider reducing computational costs either by resorting to a surrogate model of the process considered (Kennedy and O'Hagan 2001; Higdon et al. 2008; Cui et al. 2011; Laloy et al. 2013) or by developing a strategy to reduce the dimensionality of the stochastic inverse problem. Here we focus on the latter strategy and explore its effectiveness by way of a suite of computational examples.

5. Model Selection Criterion

The strong heregeoneity of the domain we consider leads to a KLE characterized by a high number of terms. Inferring the posterior joint pdf (10) through MCMC for these types of high-dimensional problems is practically unaffordable. It is then desirable to further reduce the dimensionality of the inverse problem before running the MCMC sampler. We propose doing so via the use of a model selection criterion. As an example, here we rely on the Kashyap information criterion, *KIC* (Kashyap, 1982), other alternatives (e.g., *AIC* (Akaike 1974), *AICc* (Hurvich and Tsai 1989) or *BIC* (Schwarz 1978) being fully compatible with our procedure.

The expression for *KIC* is derived from the Bayesian Model Evidence (BME) defined as

$$p\left(\boldsymbol{m}\middle|M_k\right)=\int_{KL_k} p\left(\boldsymbol{m}\middle|M_k,\boldsymbol{\xi}\right)p\left(\boldsymbol{\xi}\middle|M_k\right)d\boldsymbol{\xi} \tag{13}$$

where $\left\{M_k, k=1,...,N_k\right\}$ is a set of competing alternative models and M_k depends on KL_k quantities collected in vector $\boldsymbol{\xi}$. BME (13) is a metric quantifying how likely model M_k is, given the data $\boldsymbol{m}$. The competive models we consider in our framework are all the possible KLEs.

The analytical evaluation of the integral in (13) is not straightforward, especially for high-dimensional parameter spaces. An approximate form of (13) can be obtained by employing the Laplace approximation. The latter assumes that the posterior distribution of the parameters in $\boldsymbol{\xi}$ is Gaussian and highly peaked around its local maximum a posteriori (MAP) estimate $\boldsymbol{\xi}^{MAP}$. Expressing $p\left(\boldsymbol{m}\middle|M_k\right)$ through a Taylor series expansion centered at the MAP, retaining terms up to second-order and taking the exponential of the resulting expansion yields (see Schoniger et al. 2014)

$$p\left(\boldsymbol{m}\middle|M_k\right)= p\left(\boldsymbol{\xi}^{MAP}\middle|M_k\right)p\left(\boldsymbol{m}\middle|M_k,\boldsymbol{\xi}^{MAP}\right)\left(2\pi\right)^{K/2}\left|\boldsymbol{H}\right|^{-1/2} \tag{14}$$

where $\boldsymbol{H}$ is the Hessian matrix evaluated at the MAP, usually approximated by the Fisher information matrix $\mathbf{F}$. One then defines *KIC* as

$$KIC_k = -2\ln\left(p\left(\boldsymbol{m}\middle|\boldsymbol{\xi}^{MAP}, M_k\right)\right) - 2\ln\left(p\left(\boldsymbol{\xi}\middle|M_k\right)\right) - K\ln(2\pi) + \ln\left(|\mathbf{F}|\right) \tag{15}$$

Note that $p\left(\boldsymbol{\xi}\middle|M_k\right)$ corresponds to the prior assigned to the KL terms denoted in (10) by $p(\boldsymbol{\xi})$ and $p\left(\boldsymbol{m}\middle|\boldsymbol{\xi}, M_k\right)$ is the likelihood denoted $p\left(\boldsymbol{m}\middle|\boldsymbol{\xi}\right)$ in (10).

6. Strategy for Dimensionality Reduction of the Inverse Problem

As stated in Section 4, the approach we employ to reduce the dimensionality of the inverse problem relies on representing the *Y* field via a sparse truncated KL parameterization. The strongly heterogeneous random fields we consider are characterized by a small correlation scale, relative to a characteristic length scale of the flow domain. Values of *Y* in these fields tend to alternate rapidly in space in a rough rather than a smooth manner and treating them through KLE still requires considering a notably high-dimensional parameter space to capture the major details of the underlying field. This element constitutes a critical challenge and tends to hamper the effectiveness of characterizing the *Y* field through Bayesian inference approaches based on MCMC samplers. To alleviate this difficulty, we propose a strategy to further reduce the dimensionality of the parameterization of the problem. We construct models with different degrees of complexity through sparse KLE and evaluate their performance in the presence of available observations. We associate the degree of complexity of a model with the number of parameters which are retained in (7). Our model selection strategy is driven by available information content and is based on the use of model selection criteria of the kind illustrated in Section 5 which we employ to guide the identification of the eigenmodes (i.e., the number of parameters) of the sparse KLE which are most influential to the interpretation of the observed data.

We start by recasting the truncated KLE (7) as

$$Y(\boldsymbol{x},\omega) \approx \mu_Y + \sum_{i=1}^{K} \theta_i(\omega)\varphi_i(\boldsymbol{x}) \tag{16}$$

where, $\theta_i = \sqrt{\lambda_i}\xi_i$ and the parameter prior is now defined as $\theta_i \sim N(0,\lambda_i)$. Since the set of eigenfunctions $\{\varphi_i(\boldsymbol{x})\}_{i=1}^{K}$ are orthogonal within the spatial domain D, (16) is a variance decomposition of $Y(\boldsymbol{x},\omega)$, i.e.,

$$E_D\left[\left(Y(\boldsymbol{x},\omega)-\mu_Y\right)^2\right] = \frac{1}{D}\int_D \left(Y(\boldsymbol{x},\omega)-\mu_Y\right)^2 \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{K}\theta_i^2(\omega) \tag{17}$$

Note that the spatial variance depends on ω, i.e., on the random realization (or draw) considered. Suppose that the MAP estimate θ^{MAP} is considered. Then, (17) indicates that $\left(\theta_i^{MAP}\right)^2$ is a measure of the contribution of the i^{th} eigenmode to the spatial variance of the stochastic field. The key idea underlying the approach is that eigenmodes with negligible contribution to (17) can be discarded from the expansion (16) so that dimensionality reduction of the inverse problem can be achieved. We do so according to the procedure detailed in the following where we assume, for the sake of simplicity, that the posterior pdf (10) is unimodal.

1. Start by retaining the first K eigenmodes of the covariance function that capture most of the energy of the stochastic process. As an example, in our demonstration examples we select

$$\sum_{i=1}^{K}\lambda_i \Big/ D\sigma^2 \geq 0.90 \tag{18}$$

2. Find the maximum a posteriori estimate, $\theta^{MAP} = \arg\max_{\theta}\left(p\left(\theta|\boldsymbol{m}\right)\right)$; here, we do so by relying on the Levenberg-Marquardt (LM; Levenberg 1944; Marquardt 1963) algorithm.
3. Compute the value of a given model selection criterion. As a reference metric, we consider the *KIC* (Kashyap 1982) criterion (15) reformulated here as, (ANIS: what is

the difference between K within the parenthesis and "K at the subscript"? Are they the same thing?)

$$KIC_K = -2\ln\left(p\left(\boldsymbol{m}\middle|\boldsymbol{\theta}^{\boldsymbol{MAP}},K\right)\right)-2\ln\left(p\left(\boldsymbol{\theta}\middle|K\right)\right)-K\ln\left(2\pi\right)+\ln\left(\left|\mathbf{F}\right|\right) \tag{19}$$

Here, K indicates the number of terms retained in the KL expansion, $p\left(\boldsymbol{m}\middle|\boldsymbol{\theta}^{\boldsymbol{MAP}},K\right)$ is the likelihood function evaluated at the MAP estimate (Schöniger et al. 2014); $p\left(\boldsymbol{\theta}\middle|K\right)$ is the prior pdf of the current K KL-terms (recall that $p\left(\theta_i\middle|K\right)\sim N\left(0,\lambda_i\right)$); $\left|\mathbf{F}\right|$ is the determinant of the so-called Fisher information matrix evaluated at the MAP.

4. Compute the contribution of the i^{th} eigenmode to the spatial variance of the stochastic field, as quantified by the partial variance $\left(\theta_i^{\boldsymbol{MAP}}\right)^2$ for i=1,…,K.

5. Sort the eigenmodes $\left(\lambda_i,\varphi_i\left(\boldsymbol{x}\right)\right)$ according to their partial variance (from largest to smallest $\left(\theta_i^{\boldsymbol{MAP}}\right)^2$; see (17)).

6. Keep the K^{new} most significant eigenmodes, such that

$$\sum_{i=1}^{K^{new}}\left(\theta_i^{MAP}\right)^2\Big/\sum_{i=1}^{K}\left(\theta_i^{MAP}\right)^2\geq 0.90 \tag{20}$$

7. If K^{new}=1, then go to step 8 of the procedure; otherwise, set $K=K^{new}$, construct a new sparse KLE and go to step 2.

8. Finally, set $K^{opt}=\arg\min_K\left(KIC_K\right)$ and use DREAM$_{(ZS)}$ to sample the sparse KLE coefficients according to the target pdf $p\left(\theta\middle|\boldsymbol{m}\right)$.

Hence, step 8 yields the optimal sparse KLE, analyzed on the basis of the chosen information criterion (19). The Bayesian inference of the values of the reduced subset of parameters $\left\{\theta_i\right\}_{i=1}^{K^{opt}}$ is then performed with the MCMC DREAM$_{(ZS)}$ sampler.

Note that while we assume here that the target pdf (10) is unimodal, the procedure can be extended to the case of multimodal distributions by searching in step 2 for all optimum values obtained using multiple starting points in the LM algorithm.

7. Results and discussion

7.1 Setting of the inverse problem

We analyze and exemplify the performance of our approach upon relying on a set of computational studies performed on synthetic systems. We consider a two-dimensional square domain of side L = 10 m discretized with a mesh formed by 10,000 uniform square elements. The steady-state flow problem described by (1) is solved under permeameter-like boundary conditions corresponding to uniform (in the average) groundwater flow driven by a given head drop. As a test bed for our approach, and following the discussion of Section 3, we consider the exponential covariance function (6) with a given correlation length η/L = 0.1 and a variance $\sigma^2 = 1$. An unconditional realization of the heterogeneous Y field which we consider as reference is generated using the KLE with 400 terms. Figure 1 depicts the cumulative sum of the normalized eigenvalues (9) for the setting considered. These results suggest that a number of terms $K \approx 150$ is required for the KLE to capture about 90% of the system variance.

The steady-state forward flow problem is then solved for the generated reference Y field. Values of Y and hydraulic head are jointly sampled at 25 diverse locations randomly selected in the system and constitute the entries of the vector $\boldsymbol{m}$ of observation data. We assume that both head and Y measurements are noisy. Measurement errors are considered to be uncorrelated in space and are modeled as zero-mean Gaussian random variables, characterized by known standard deviations, denoted as σ_h and σ_Y, respectively for head and

Y data. Figure 2 depicts the reference *Y* field and the 25 locations at which observations of both *Y* and hydraulic head are collected in our example.

Following Bayes' theorem, the posterior pdf of the KLE modes is given by

$$p\left(\boldsymbol{\theta}\middle|\boldsymbol{m},K,\sigma_h,\sigma_Y,\mathbf{C}\right)\propto exp\left(-\frac{SS_1\left(\boldsymbol{\theta}\right)}{2\sigma_h^2}-\frac{SS_2\left(\boldsymbol{\theta}\right)}{2\sigma_Y^2}\right)exp\left(\frac{-1}{2}\boldsymbol{\theta}^T\mathbf{C}^{-1}\boldsymbol{\theta}\right) \tag{21}$$

where *T* is transpose and **C** is the covariance matrix defined by

$$\mathbf{C}=\begin{bmatrix}\lambda_1 & 0 & \cdots\\ 0 & \ddots & \vdots\\ 0 & \cdots & \lambda_K\end{bmatrix} \tag{22}$$

Here, $SS_1\left(\boldsymbol{\theta}\right)$ and $SS_2\left(\boldsymbol{\theta}\right)$ respectively are the sum of squared differences between observed and modeled (relying on *K* modes of the KLE) head and *Y* values. Measurement error standard deviation of pressure heads is set to $\sigma_h=0.05$ m, which corresponds to 5% of the largest head variation $\left(h_{\max}-h_{\min}\right)$ in the domain. Two scenarios corresponding to different values of standard deviation of measurement errors of *Y* are investigated, i.e., σ_Y = 0.1 and 0.5, respectively corresponding to 2% and 10% of the largest *Y* variation $\left(Y_{\max}-Y_{\min}\right)$ across the domain.

Consistent with the assumptions in the approach underlying (18), the information matrix **F** embedded in *KIC* (19) is rendered by (Schöniger et al. 2014)

$$\mathbf{F}=\mathbf{J}^T\boldsymbol{\Sigma}^{-1}\ \mathbf{J}+\mathbf{C}^{-1} \tag{23}$$

where **J** is the Jacobian matrix evaluated at MAP and $\boldsymbol{\Sigma}$ the covariance matrix defined as

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_Y^2\mathbf{I}_{N_{obs}/2} & 0\\ 0 & \sigma_h^2\mathbf{I}_{N_{obs}/2}\end{bmatrix}; \tag{24}$$

N_{obs} being the number of data collected in the vector $\boldsymbol{m}$ and $\mathbf{I}_{N_{obs}/2}$ the identity matrix of size *Nobs*/2.

We remark that Bayesian inversion with MCMC using the KLE of the Y field associated with K = 150, which allows capturing approximately 90% of the variance associated with the postulated exponential covariance function (6), was unaffordable due to the large number of parameters. The following section is devoted to the illustration of our application of the dimensionality reduction strategy described in Section 5.

7.2 KLE with dimensionality reduction

We apply the model reduction strategy described in Section 6 starting from the KLE associated with K = 150. The components of the MAP vector θ^{MAP} are estimated through the LM algorithm and the corresponding value of *KIC* (19) is computed following steps 1-4 of the algorithm described in Section 6. The algorithm is continued until only one term remains in the sparse KLE. This screening phase required about 370 model calls and is computationally cheap as compared to the the cost required by MCMC samplers (around 50,000 model calls).

Figure 3 depicts the dependence of *KIC* on the number of modes ($1 \leq K \leq 150$) retained in the sparse KLE and resulting from the application of the reduction procedure decribed in Section 6. This figure indicates that *KIC* identifies a minimum corresponding to the use of solely 19, or 12 components of the sparse KLE, respectively for σ_Y = 0.1 and 0.5. In other words, the information content embedded in the available noisy measurements allows identifying a sparse KLE representation of the Y field based on a reduced number of components, i.e., K = 19, or 12 in the cases analyzed. This result is consistent with the general idea that a reduced number of parameters is required to interpret data associated with large measurement errors. We note that we obtain results of similar quality by relying also on diverse quantities, such as AIC (Akaike 1974) or BIC (Schwarz 1978) criteria (not shown). When sorted in order of importance, the modes retained at the optimum correspond to the components identified by the sets of indices {i = 2, 17, 21, 49, 7, 38, 69, 8, 28, 79, 41, 33, 36,

20, 40, 80, 78, 13, 10} or {i = 2, 8, 36, 49, 17, 30, 21, 79, 38, 122, 129, 6}, respectively for σ_Y = 0.1 and 0.5. We recall here that modes are selected and ranked according to their relevance (see (17) and step 5 in the reduction algorithm).

Finally, the resulting Y field parameterizations are employed to appraise the posterior pdf (21) through DREAM$_{(ZS)}$. Figure 4 depicts the inferred posterior marginal pdfs of the first three KL modes identified by the set of indices listed above and resulting from stochastic model calibration via MCMC for the two scenarios examined. These results reveal that the mode values are appropriately estimated. Their associated posterior pdfs are unimodal, with an approximately symmetric shape, and encompass a narrow range of values for both values of σ_Y considered. Results of similar quality are obtained for the remaining modes retained in these sets (not shown).

Figure 5 depicts the results of the MCMC-based inversion evaluated at the measurement locations for h and Y and for both values of σ_Y tested. The 95% uncertainty bounds (corresponding to the 97.5 and 2.5 percentiles of the distributions) representing parametric uncertainty (narrow bounds in the figure) are depicted in Figure 5 together with the total predictive uncertainty (wide bounds in the figure), the latter taking into account parametric uncertainty as well as measurement errors. The results of Figure 5 suggest that virtually all observations are comprised within the 95% total uncertainty range for both values of σ_Y. As expected, the total uncertainty characterizing Y estimates tends to increase with σ_Y. The parametric uncertainty is slightly larger for σ_Y = 0.1 than for σ_Y = 0.5, respectively involving 19 and 12 modes at the optimum.

Figure 6a, b depict the MAP estimate of the spatial field Y, respectively for σ_Y = 0.1, and 0.5. Figure 6c, d depict the spatial distribution of the width of the 95% total uncertainty ranges of h, respectively for σ_Y = 0.1, and 0.5. The corresponding graphical depiction for the

width of the 95% uncertainty ranges of Y is shown in Figures 6e, f. Direct comparison of Figures 6a, b and Figure 2 suggests that the identified (optimum) sparse KLEs yield a good MAP approximation of the reference log-transmissivity field, with a good quality representation of the spatial pattern of poorly and highly conducive regions, for both cases. It is nevertheless noted that, even as the MAP estimate can be deemed satisfactory, the predictive total uncertainty (Figure 6c-f) associated with the stochastic field tends still to be large at locations far from measurements. This feature is especially evident for $\sigma_Y = 0.5$.

7.3 Predictive Performance

Figure 5 suggests that the calibrated models provide a satisfactory representation of the observations in a probabilistic sense. We now analyze their predictive performance at diverse locations in the domain. The reference values at unsampled locations can be compared against the corresponding MCMC predictive distributions of $h(\boldsymbol{x})$ and $Y(\boldsymbol{x})$. The estimated Cumulative Distribution Functions (CDFs) obtained for h and Y are respectively depicted in Figures 7 and 8 together with the corresponding reference value for $\sigma_Y = 0.1, 0.5$. Only a set of selected locations in the domain are displayed, as representative of the range of results obtained in our simulations. It can be noted that at some locations the reference value is comprised within the range of values associated with non-negligible probability for the two CDFs depicted. Otherwise, there are locations at which this behavior can be observed for only one of the two posterior CDFs, which is most frequently linked to the largest variance of the measurement errors. Nonetheless, there are some locations (far from measurements) where the reference values are not captured by either of the CDFs obtained from our inversion. Hence, the parameterization strategy based on the identification of a reduced dimensionality KLE may lead to collections of solutions which do not encompass the reference solution at some unsampled locations (far from measurements). To improve the quality of the estimation, one can, for instance, increase the number of measurements and/or the threshold for the

selection of eigenmodes in the MAP to yield an augmented number of KL eigenmodes, thus contributing to improve the quality of the inverse solutions (as compared to the reference solution).

8. Conclusions

We develop an operational strategy to obtain computationally affordable and Bayesian estimates of satisfactory quality of heterogeneous transmissivity fields in the presence of sampled data available at a set of locations in an aquifer. We do so by relying on a scheme based on modeling the (natural) logarithm of transmissivity as a stochastic Gaussian process which is parameterized through a truncated KLE. We consider strongly heterogeneous transmissivity fields, such as those characterized by short-range (with respect to the domain size) correlation, for which Bayesian inference becomes highly challenging and computationally demanding due to the large number of terms which are required to be retained in the KLE.

Our strategy starts from a highly-parameterized field and yields a set of sparse KLEs with reduced dimensionality, the MAP estimate of the eigenmodes in each sparse KLE being obtained through inverse modeling of flow against noisy data. Selection of the optimal number of modes to be retained in the expansion is driven by a model selection criterium, which is informed by available observations. The posterior statistical distribution of the corresponding eigenmodes is then obtained upon relying on the DREAM$_{(ZS)}$ MCMC sampler developed by Laloy and Vrugt (2012).

The approach is illustrated by relying on a suite of computational examples where noisy transmissivity and head values are sampled from a given transmissivity field. The new methodology yields a satisfactory inversion of the stochastic field with a good representation of the observations in a probabilistic sense. At some unsampled locations (far from measurements), the collection of estimated solutions may not encompass the reference values.

The quality of the estimation could be improved for instance by increasing the number of measurements and/or the threshold for the selection of KL eigenmodes in the MAP.

Acknowledgements

The authors are grateful to the French National Research Agency who funded this work through the program AAP Blanc - SIMI 6 project RESAIN (n° ANR-12-BS06-0010-02).

AG acknowledges funding from the European Union's Horizon 2020 Research and Innovation programme in the context of the Water JPI (WATERWORKS2014 ERA-NET cofunded program; Project "WatEr NEEDs, availability, quality and sustainability" WE-NEED).

References

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control AC-19: 716-723. doi: 10.1109/TAC.1974.1100705

Chen X, Murakami H, Hahn MS, Hammond GE, Rockhold ML, Zachara JM, Rubin Y (2012) Three-dimensional Bayesian geostatistical aquifer characterization at the Hanford 300 Area using tracer test data. Water Resour Res 48 W06501 doi: 10.1029/2011WR010675

Cui T, Fox C, O'Sullivan MJ (2011) Bayesian calibration of a large scale geothermal reservoir model by a new adaptive delayed acceptance metropolis hastings algorithm. Water Resour Res 47 W10521. doi:10.1029/2010WR010352.

Dagan G (1989), Flow and Transport in Porous Formations, Springer-Verlag, New York.

Das NN, Mohanty BP, Efendiev Y (2010) Characterization of effective saturated hydraulic conductivity in an agricultural field using Karhunen-Loève expansion with the Markov chain Monte Carlo technique. Water Resour Res 46 W06521. doi:10.1029/2008WR007100.

Efendiev Y, Hou TY, Luo W (2006) Preconditioning Markov chain Monte Carlo simulations using coarse-scale models. SIAM Journal on Scientific Computing 28:776-803. doi: 10.1137/050628568

ANIS: CHECK THE REFERENCE, especially the number of pages. Many thanks

Gneiting T, Genton MG, Guttorp P (2007) Geostatistical space-time models, stationarity, separability, and full symmetry, Monographs On Statistics and Applied Probability 107: 151.

Green PJ, Mira A (2001) Delayed rejection in reversible jump Metropolis-Hastings. Biometrika 88: 1035–1053. doi: 10.1093/biomet/88.4.1035

Haario H, Saksman E, Tamminen J (2001) An adaptive Metropolis algorithm. Bernouilli 7, 2:223-242.

Higdon D, Gattiker J, Williams B, Rightley M (2008) Computer model calibration using high-dimensional output. J Am Stat Assoc. 103:570–583. doi: 10.1198/016214507000000888

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small sample, Biometrika 76(2): 297-307.

Huard D, Mailhot A, Duchesne S (2010) Bayesian estimation of intensity–duration–frequency curves and of the return period associated to a given rainfall event, Stoch Environ Res Risk Assess. 24(3): 337-347. doi: 10.1007/s00477-009-0323-1.

Kashyap RL (1982) Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Trans Pattern Anal Mach Intell 4 (2):99-104. doi: 10.1109/TPAMI.1982.4767213

Keating EH, Doherty J, Vrugt JA, Kang Q (2010) Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. Water Resour Res 46, W10517. doi: 10.1029/2009WR008584

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc 63 (B): 425–464. doi: 10.1111/1467-9868.00294

Laloy E, Rogiers B, Vrugt JA, Mallants D and Jacques D (2013) Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. Water Resour Res 49:2664-2682. doi: 10.1002/wrcr.20226

Laloy E, Vrugt JA (2012) High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing. Water Resour Res 48, W01526, doi: 10.1029/2011WR010608

Levenberg K (1944) A method for the solution of certain non-linear problems in least squares. The Quarterly of Applied Mathematics 2:164-168.

Li W, Cirpka OA (2006) Efficient geostatistical inverse methods for structured and unstructured grids. Water Resources Research 42, W06402. doi: 10.1029/2005wr004668

Lin G, Tartakovsky AM, Tartakovsky DM (2010) Uncertainty quantification via random domain decomposition and probabilistic collocation on sparse grids. J Comput Phys 229: 6995-7012. doi:10.1016/j.jcp.2010.05.036

Loeve M (1977) Probability Theory, fourth ed., Springer, New York.

Mara TA, Fajraoui N, Younes A, Delay F (2015) Inversion and uncertainty of highly parameterized models in a Bayesian framework by sampling the maximal conditional posterior distribution of parameters. Advances in Water Resources 76: 1- 10. doi: 10.1016/j.advwatres.2014.11.013

Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. SIAM Journal on Applied Mathematics 11, 431–441. doi: 10.1137/0111030

Marzouk Y, Najm HN (2009) Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. Journal of Computational Physics 228:1862-1902. doi: 10.1016/j.jcp.2008.11.024

Mercer J (1909) Functions of positive and negative type and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society; 209 :415-446.

Murakami H, Chen X, Hahn MS, Liu Y, Rockhold ML, Vermeul VR, Zachara JM, Rubin Y (2010) Bayesian approach for three-dimensional aquifer characterization at the Hanford 300 Area. Hydrol Eart Syst Sci 14:1989-2001. doi:10.5194/hess-14-1989-2010

Over MW, Chen X, Yang Y, Rubin Y (2013) A strategy for improved computational efficiency of the method of anchored distributions. Water Resour Res 49:1-19. doi: 10.1002/wrcr.20182

Phoon KK, Huang SP, Quek ST (2002) Implementation of karhunen-loeve expansion for simulation using a wavelet-galerkin scheme. Probabilistic Engineering Mechanics 17:293–303. doi:10.1016/S0266-8920(02)00013-9.

Ray J, McKenna SA, van Bloemen Waanders B, Marzouk YM (2012) Bayesian reconstruction of binary media with unresolved fine-scale spatial structures. Advances in Water Resources 44 (2012) 1-19. doi: 10.1016/j.advwatres.2012.04.009

Rubin Y, Chen X, Murakami H, Hahn M (2010) A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields. Water Resour Res 46, W10523. doi:10.1029/2009WR008799

Schwarz GE (1978) Estimating the dimension of a model, Annals of Statistics 6 (2):461–464. doi:10.1214/aos/1176344136, MR 468014

Schöniger A, Wöhling T, Samaniego L, Nowak W (2014) Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. Water Resour Res. 50:9484-9513. doi:10.1002/2014WR016062.

Schoups G, Vrugt JA (2010) A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors, Water Resour Res 46, W10531. doi:1029/2009WR008933.

Schwarz G (1978) Estimating the dimension of a model, Ann. Stat. 6(2): 461-464.

Shi X, Ye M, Finsterle S, Wu J (2012) Comparing nonlinear regression and Markov chain Monte Carlo methods for assessment of predictive uncertainty in vadose zone modeling. Vadose Zone J 11 (4). doi: 10.2136/vzj2011.0147

Spanos Pol D, Beer M, Red-Horse J (2007) Karhunen-Loève expansion of stochastic processes with a modified exponential covariance kernel. Journal of Engineering Mechanics 133: 773-779.

Su C-H, Lucor D (2006) Covariance kernel representations of multidimensional second-order stochastic processes. Journal of Computational Physics 217: 82-99.

Tartakovsky DM, Nowak W, Bolster D (2013) Introduction to the special issue on uncertainty quantification and risk assessment Adv Water Resour 36:1-2. doi: 10.1016/j.advwatres.2011.12.010

Tartakovsky DM (2013) Assessment and management of risk in subsurface hydrology: A review and perspective. Adv Water Resour 51:247-260. doi:10.1016/j.advwatres.2012.04.007

ter Braak C, Vrugt J (2008) Differential Evolution Markov Chain with snooker updater and fewer chains. Statistics and Computing 18(4): 435-446 doi: 10.1007/s11222-008-9104-9

Tsantili IC, Hristopulos DT (2016) Karhunen-Loève expansion of spartan spatial random fields. Probabilistic Engineering Mechanics 43:132-147.

Vrugt JA, Bouten W (2002) Validity of First-Order Approximations to Describe Parameter Uncertainty in Soil Hydrologic Models. Soil Sci Soc Am J 66:1740–1751. doi: 10.2136/sssaj2002.1740

Vrugt JA, Gupta HV, Bouten W, Sorooshian S (2003) A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour Res 39(8)1201. doi: 10.1029/2002WR001642

Vrugt JA, ter Braak CJF, Clark MP, Hyman JM, Robinson BA (2008) Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour Res 44 W00B09. doi:10.1029/2007WR006720

Vrugt JA, ter Braak CJF, Diks CGH, Higdon D, Robinson BA, Hyman JM (2009a) Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int J Nonlinear Sci Numer Simul 10(3): 273–290. doi: 10.1515/IJNSNS.2009.10.3.273

Vrugt JA, ter Braak CJF, Gupta HV, Robinson BA (2009b) Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stoch Environ Res Risk Assess. 23(7): 1011-1026. doi:10.1007/s00477-008-0274-y

Younes A, Ackerer P, Delay F (2010) Mixed finite element for solving 2D diffusion-type equations. Reviews of Geophysics 48 RG1004. doi: 10.1029/2008RG000277

Zanini A, Kitanidis PK (2009) Geostatistical inversing for large-contrast transmissivity fields. Stoch Environ Res Risk Assess. 23:565–577. doi: 10.1007/s00477-008-0241-7

Zhang D (2002) Stochastic methods for flow in porous media, coping with uncertainties, Academic Press, San Diego.

Zhang D, Lu Z (2004) An efficient, higher-order perturbation approach for flow in randomly heterogeneous porous media via Karhunen-Loeve decomposition. J Comput Phys 194:773–794. doi : 10.1016/j.jcp.2003.09.015

Zheng Y, Han F (2016) Markov Chain Monte Carlo (MCMC) uncertainty analysis for watershed water quality modeling and management. Stoch Environ Res Risk Assess 30(1):293–308. doi: 10.1007/s00477-015-1091-8

Figure Captions

Figure 1: Cumulative sum of the normalized eigenvalues (see (9)) for the exponential covariance with $\eta / L = 0.1$ and variance $\sigma^2 = 1$.

Figure 2. Reference spatial field of the log-transmissivity field, $Y(\boldsymbol{x})$. Crosses indicate locations where head and *Y* values are jointly sampled.

Figure 3. Selection of the optimal number of modes, K^{opt}, based on the *KIC* model selection criterion (16) for the values of standard deviation of data measurement errors: (left) $\sigma_h = 0.05$, $\sigma_Y = 0.1$ and (right) $\sigma_h = 0.05$, $\sigma_Y = 0.5$.

Figure 4. Inferred posterior probability distribution of selected KL eigenmodes after statistical calibration with MCMC for the values of standard deviation of data measurement errors: (left column) $\sigma_h = 0.05$, $\sigma_Y = 0.1$ and (right column) $\sigma_h = 0.05$, $\sigma_Y = 0.5$.

Figure 5. MCMC predictive uncertainty of the statistically calibrated reduced models. First row: data are corrupted through Gaussian errors with standard deviation $\sigma_h = 0.05$ (for heads) and $\sigma_Y = 0.1$ (for log-transmissivity). Second row: data are corrupted with Gaussian errors with $\sigma_h = 0.05$ (for heads) and $\sigma_Y = 0.5$ (for log-transmissivity).

Figure 6. Results of the sparse KLE inversion with DREAM$_{(ZS)}$ MCMC. Data are characterized by (left column) $\sigma_Y = 0.1$ or (right column) $\sigma_Y = 0.5$. First row: MAP estimate of the *Y* field. The last two rows include the width of the 95% total predictive uncertainty range for (c, d) pressure head and (e, f) log-transmissivity.

Figure 7. Comparison between cumulative distribution functions of pressure heads at selected unsampled locations (red: 19 modes reduced sparse KLE (σ_Y = 0.1); green: 12 modes reduced sparse KLE (σ_Y = 0.5)). Blue dashed lines indicate reference values. Coordinate pairs in parenthesis correspond to the locations selected in the domain.

Figure 8. Comparison between cumulative distribution functions of log-transmissivity at selected unsampled locations (red: 19 modes reduced sparse KLE (σ_Y = 0.1); green: 12 modes reduced sparse KLE (σ_Y = 0.5)). Blue dashed lines indicate reference values. Coordinate pairs in parenthesis correspond to the locations selected in the domain.

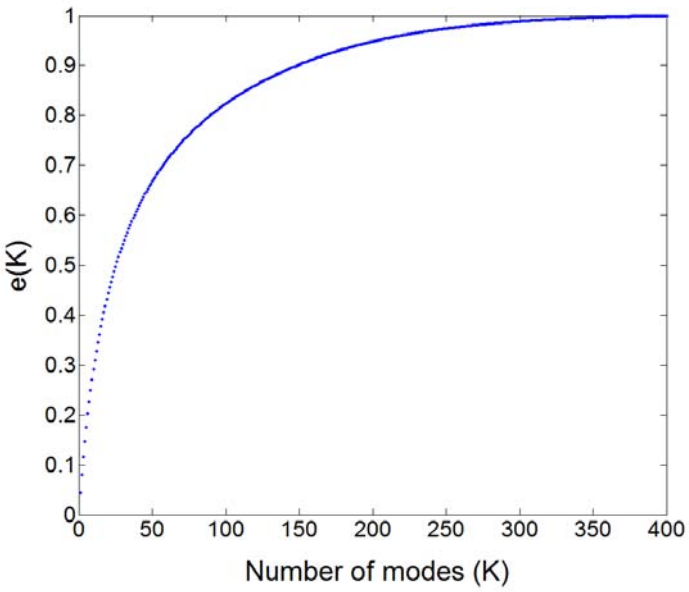

Figure 1

ANIS: the *K* should be in Italic on both axes, OK?

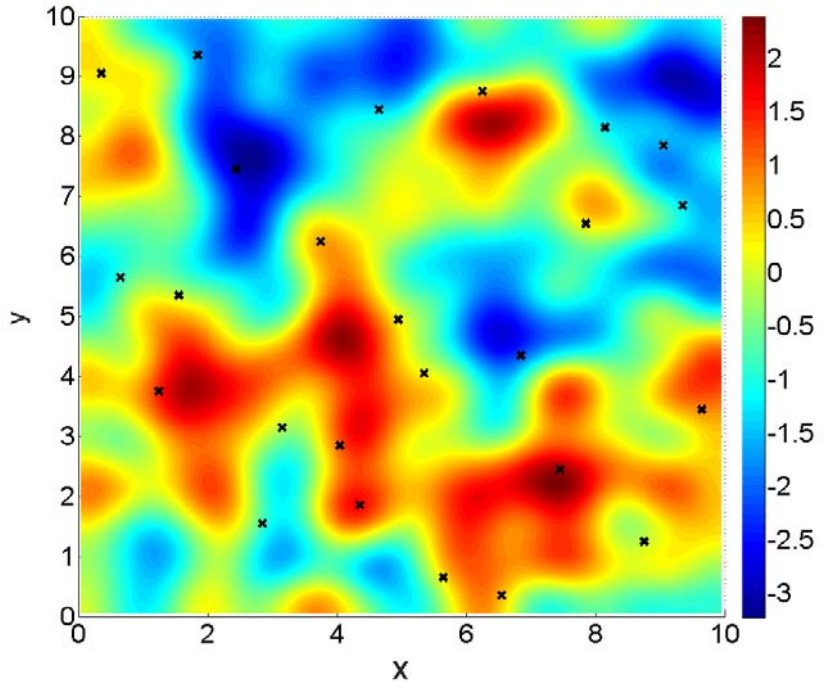

Figure 2

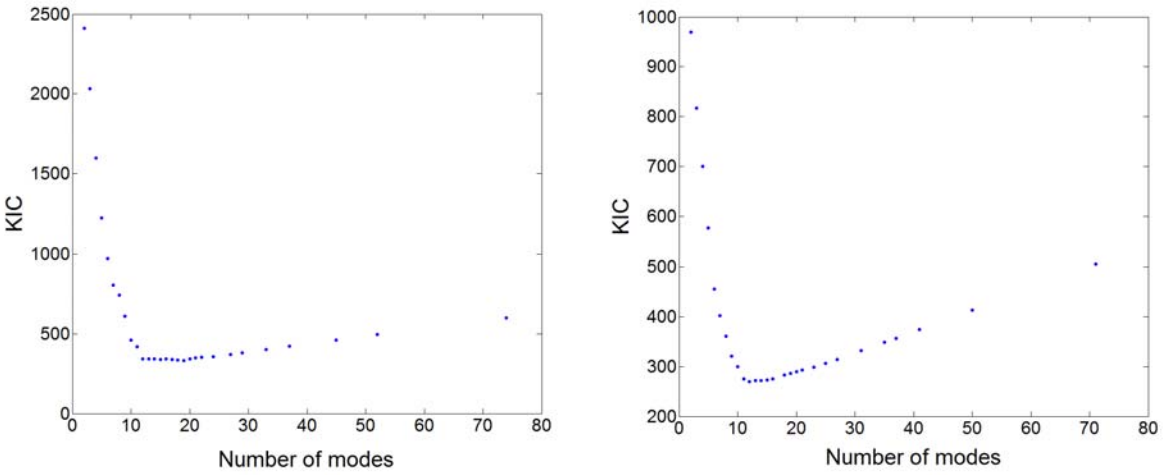

Figure 3

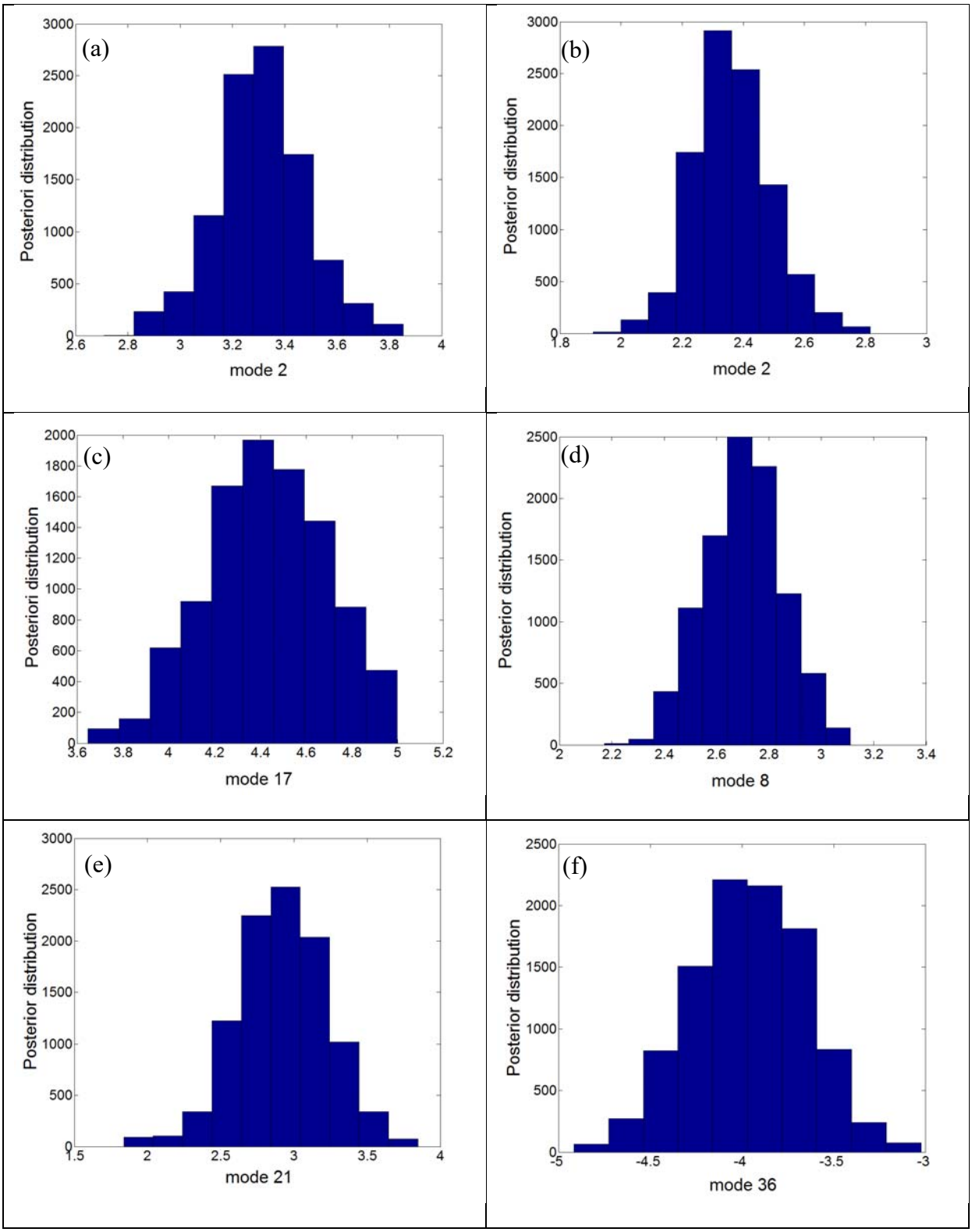
(a)
Posteriori distribution
mode 2
(b)
Posterior distribution
mode 2
(c)
Posteriori distribution
mode 17
(d)
Posterior distribution
mode 8
(e)
Posterior distribution
mode 21
(f)
Posterior distribution
mode 36

Figure 4

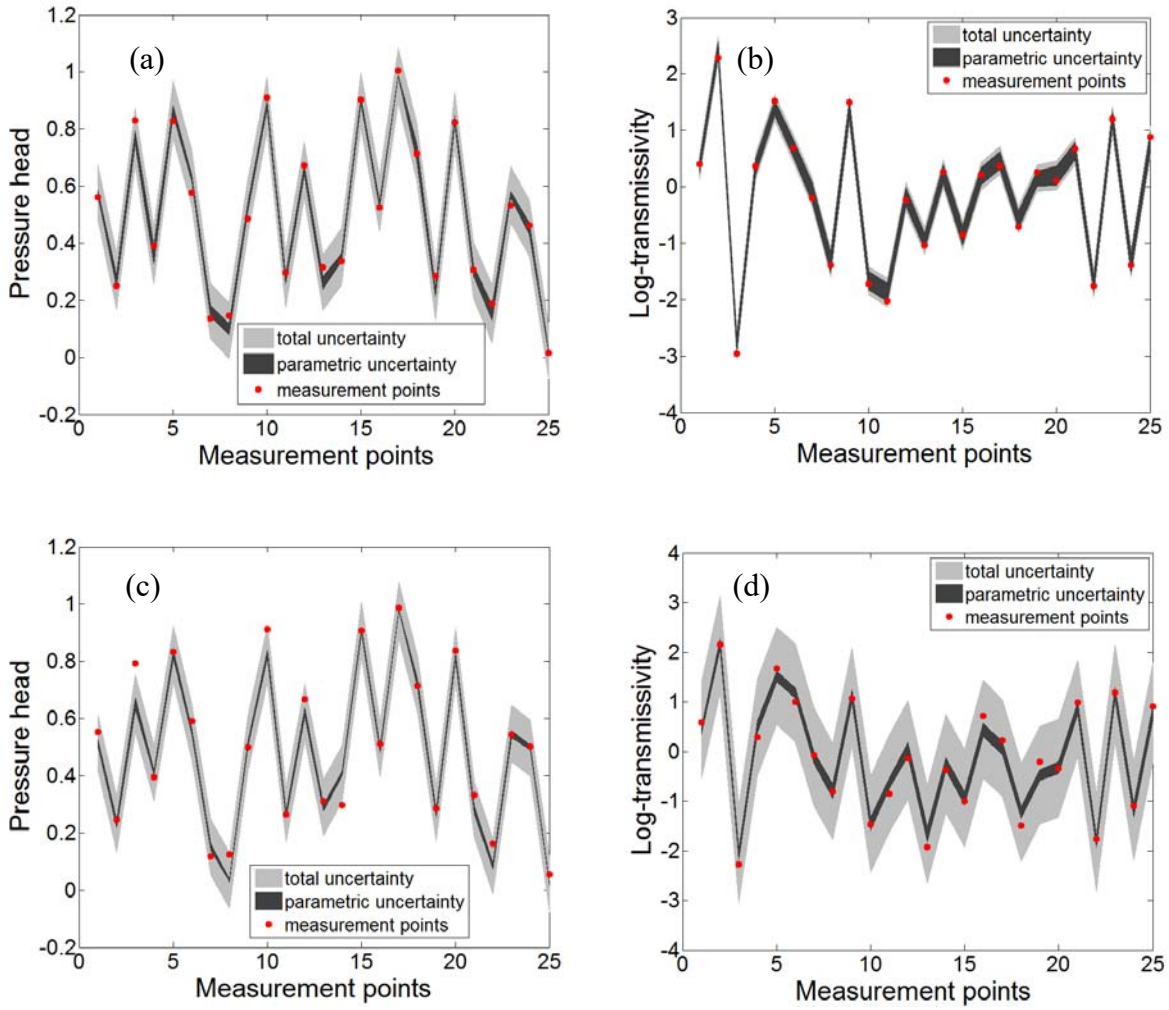
(a)
Pressure head
Measurement points
total uncertainty
parametric uncertainty
measurement points
(b)
Log-transmissivity
Measurement points
total uncertainty
parametric uncertainty
measurement points
(c)
Pressure head
Measurement points
total uncertainty
parametric uncertainty
measurement points
(d)
Log-transmissivity
Measurement points
total uncertainty
parametric uncertainty
measurement points

Figure 5

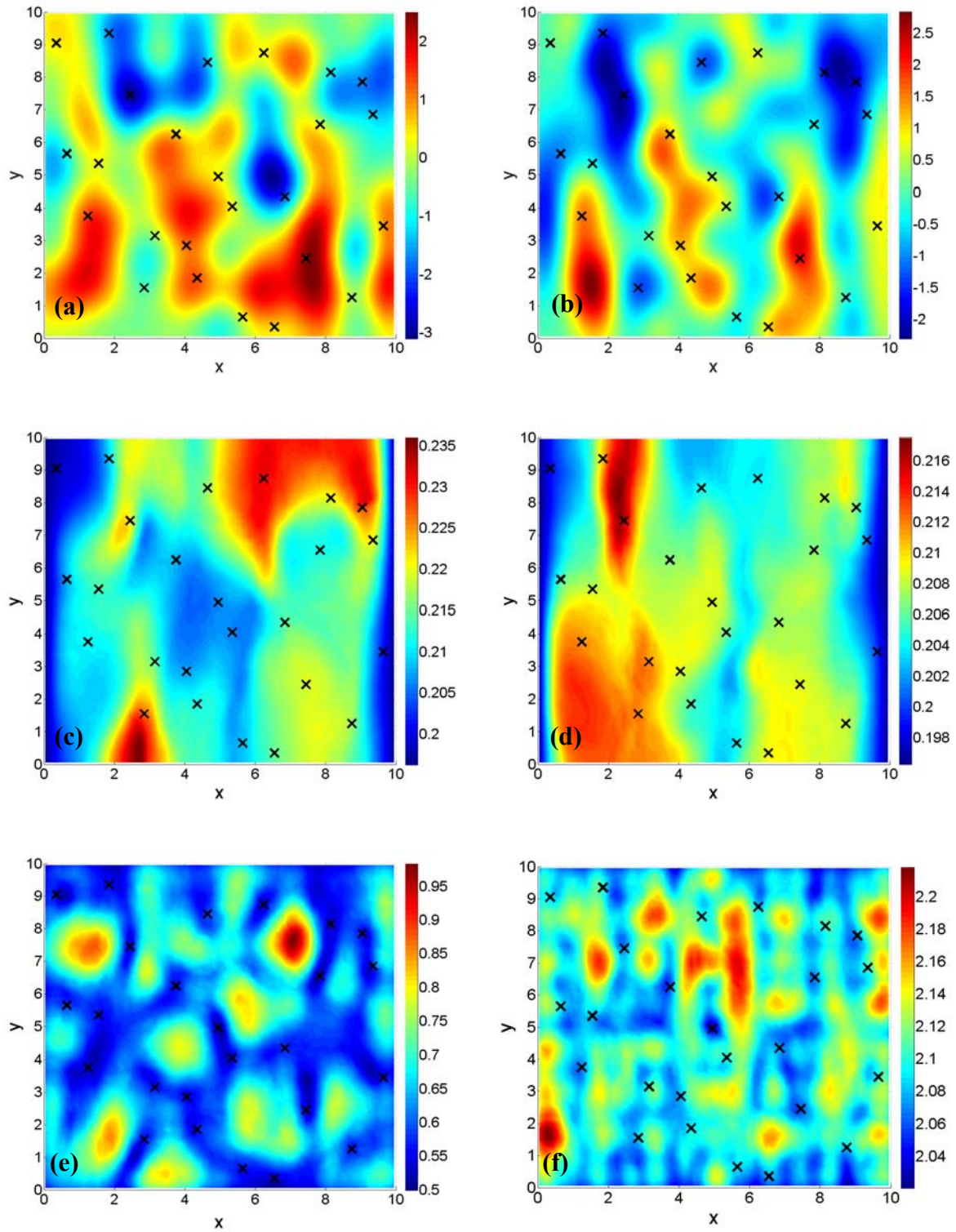

Figure 6

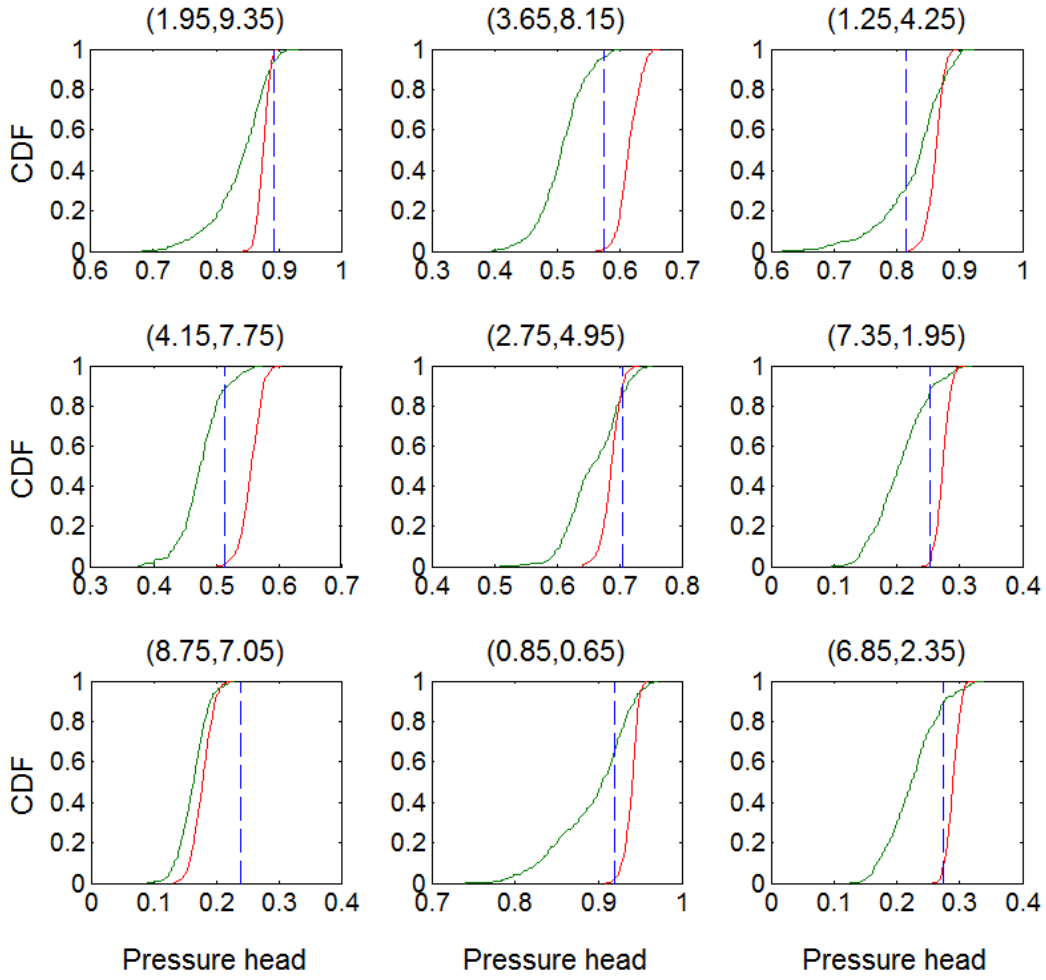
(1.95,9.35)
(3.65,8.15)
(1.25,4.25)
(4.15,7.75)
(2.75,4.95)
(7.35,1.95)
(8.75,7.05)
(0.85,0.65)
(6.85,2.35)
CDF
Pressure head

Figure 7

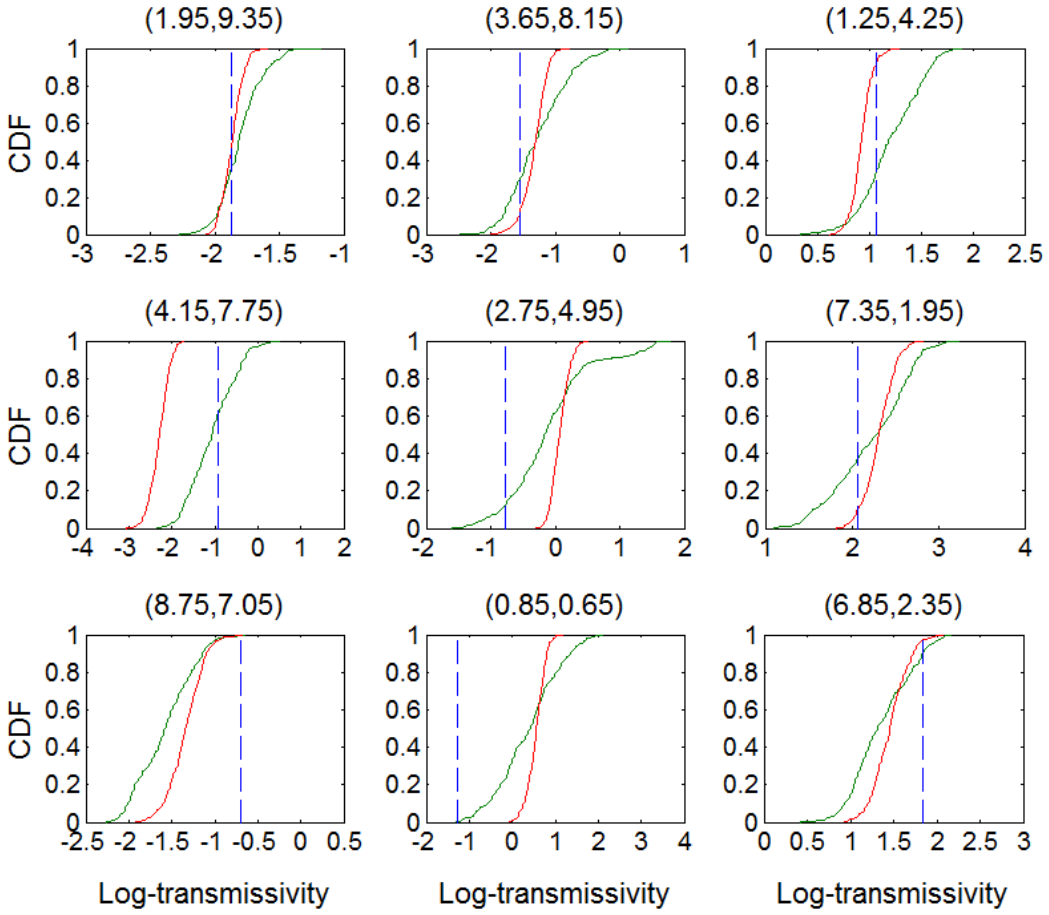
(1.95,9.35)
(3.65,8.15)
(1.25,4.25)
(4.15,7.75)
(2.75,4.95)
(7.35,1.95)
(8.75,7.05)
(0.85,0.65)
(6.85,2.35)
CDF
Log-transmissivity

Figure 8