A reduced-order model for Monte Carlo simulations of stochastic groundwater flow

Damiano Pasetto · Alberto Guadagnini · Mario Putti

Received: 26 June 2013 / Accepted: 22 November 2013 / Published online: 19 December 2013

D. Pasetto (✉)
Institut National de la Recherche Scientifique, Centre Eau Terre
Environnement (INRS-ETE), Université du Québec,
G1K 9A9 Quebec City, Quebec, Canada
e-mail: pasetto@math.unipd.it

A. Guadagnini
Dipartimento di Ingegneria Civile e Ambientale,
Politecnico di Milano, Piazza L. Da Vinci 32,
Milano, Italy
e-mail: alberto.guadagnini@polimi.it

A. Guadagnini
Department of Hydrology and Water Resources,
University of Arizona, Tucson, AZ 85721, USA

Present Address:
D. Pasetto · M. Putti
Dipartimento di Matematica, University of Padova,
Via Trieste 63, 35121 Padova, Italy

M. Putti
e-mail: mario.putti@unipd.it

1 Introduction

Modeling groundwater flow in natural aquifers requires coping with spatial heterogeneity of hydraulic properties, e.g., hydraulic conductivity and/or transmissivity. A proper characterization of these parameters is the key to, e.g., optimize water management and accurately predict transport of contaminants. A deterministic approach to the solution of the flow problem is typically based on an estimation of the spatial distribution of the hydraulic parameter fields and the subsequent solution of the equations governing heads and fluxes. A stochastic approach (e.g., [1, 2]) describes parameters, such as aquifer transmissivity, as random fields with given probability distribution and aims at rendering the probability distribution of state variables, such as hydraulic heads and fluxes. A stochastic approach is appealing when sensitivity or uncertainty analyses of hydraulic head distributions are required in the presence of incomplete knowledge on the system parameters. It also enables one to embed methodologies such as Markov chain Monte Carlo (MC) and ensemble data assimilation for aquifer characterization under uncertainty (see, e.g., [3, 4] and references therein).

Approaches that have been employed to solve the stochastic groundwater flow equation include moment differential equations (MDE) formulations, techniques based on partial differential equations satisfied by the probability density function (pdf) of the state variable of interest, and the numerical MC simulation framework. The idea underlying MDEs is the derivation of deterministic equations satisfied by the statistical (ensemble) moments of a hydraulic head from the constitutive groundwater equation and on the basis of the knowledge on the (ensemble) moments of the system parameters, e.g., the transmissivity field. The expected value and covariance of the hydraulic heads are then numerically computed by way of recursive approximations of otherwise nonlocal MDEs (see, e.g., [5, 6]). The MDE approach has been recently embedded within the context of ensemble Kalman filter-based data assimilation procedures of groundwater flow [7]. Current formulations of MDEs cannot be easily extended to provide a complete characterization of the pdf of hydraulic heads in the presence of randomly variable transmissivities, because of the relatively complex formulation and prohibitive computational effort required to compute statistical moments of order larger than two.

An alternative approach, which has been developed mainly in the context of solute transport in randomly heterogeneous groundwater velocity fields (e.g., [8–11] and references therein) relies on the development of equations governing the space-time evolution of the pdf of solute concentrations. With the exception of a few special cases, the solution of these equations typically entails a series of approximations that are still limiting the direct application of the approach to practical aquifer-scale environmental problems.

MC-based methods rely on the generation of multiple independent and identically distributed realizations of the parameter fields driving groundwater flow. The corresponding solution of the flow equation yields a collection of realizations of hydraulic heads from which ensemble statistics can be evaluated. Although the implementation of MC methods is straightforward also in the presence of highly nonlinear models, the convergence of the empirical distribution to the underlying theoretical head probability distribution is generally slow and may require a large number of computationally expensive solutions of the numerical model ([12] and references therein). For this reason, a routine application of MC simulations to real field-scale aquifer systems would significantly benefit from the development of a fast and accurate surrogate/reduced model for the computation of a large collection of hydraulic head realizations [13].

Techniques based on the polynomial chaos expansion (PCE) approximation may constitute a viable framework to obtain such surrogate system models. PCE represents hydraulic head as a series of polynomials in terms of a given set of random parameters. These polynomials are orthonormal to the joint probability measure associated with the pdf of the uncertain system parameters ([14–17], and references therein). This expansion enables the efficient computation of the moments of hydraulic head and, eventually, its complete pdf. The spatially distributed coefficients of the series are computed upon relying on the solution of the flow equation according to the Galerkin projection or the probabilistic collocation method (PCM) [18]. The number of elements to retain in the series expansion and evaluations of the original groundwater flow equation depends on the number of independent random parameters appearing in the flow equation. For this reason, the PCE is typically applied only in the presence of a limited number of random parameters, an alternative being the reliance on an approximation (e.g., truncated Karhunen–Loève expansion) of the (spatially distributed) stochastic parameters.

Surrogate models developed within the context of the Galerkin projection methods are of critical interest for the reduction of the computational cost related to the numerical evaluation of the collection of hydraulic head realizations. To achieve accurate and sizeable reduction, these surrogate models are built by projecting the original model equations onto a set of basis functions calculated from a limited number of solutions of the complete model of the system. This is typically termed *off-line* phase, the *online* phase being the ensuing application of the surrogate model to form the ensemble. The model reduction procedure is computationally advantageous when the dimension of the reduced model, i.e., the number of employed basis functions, is considerably smaller than that of the original model. Proper orthogonal decomposition (POD) [19, 20] and reduced basis (RB) [21, 22] are two model order reduction approaches that compute spatially distributed basis functions within an off-line procedure based on the snapshot technique. This technique relies on a collection of a certain number of full system model solutions, i.e., the solutions of the original model equations obtained for selected observation times and/or parameter values. POD performs a singular value analysis on the set of the snapshots. The principal components associated with the rightmost singular values constitute an optimal set of basis functions for the reproduction of the snapshots, in the sense that any other set of the same size reproduces the snapshots with a larger error. Several examples of application of POD to reduce the computational burden associated with deterministic groundwater flow problems can be found in the literature. Siade et al. [23] use POD in the context of inverse modeling to accelerate estimation of aquifer transmissivities with quasi-linearization and quadratic programming. Kaleta et al. [24] and van Doren et al. [25] develop a reduced-order model for the solution of the flow equations for reservoir simulation and the corresponding adjoint system. The principal component analysis is critical to remove redundant information when a large number of snapshots is available. The number of snapshots and principal components that one should employ to obtain a reduced model with a desired accuracy on the solution depends critically on the target model and cannot be quantified a priori. For example, Pasetto et al. [13] show that the application of the POD methodology in the presence of a stochastic and spatially distributed recharge (which constitutes an additive noise for the flow equation) is strongly affected by the variance and correlation length of the recharge term.

The RB approach (see, e.g., [22]) relies on the computation of the full system model solution only for those snapshots that maximize the amount of information to be embedded in the reduced model. In this case, the orthonormalized snapshots are directly selected as basis functions, thus circumventing the need for a principal component analysis. To determine the dimension of the reduced model, a validation set of parameter values is first considered; new snapshots are then sequentially added to the reduced model until the full model solutions employed in the validation set are reproduced within a given level of accuracy. The procedure for the selection of the snapshots is based on the so-called greedy algorithm. Given a set of basis functions, a new snapshot is computed by considering the parameter realization (selected amongst the parameter realizations forming the validation set) that maximizes the discrepancy between the full system model and the reduced model solutions. The algorithm terminates when the maximum estimated error on each hydraulic head solution of the validation set falls below a preselected tolerance. Since the error between the reduced-order and the full system model solutions cannot be explicitly computed (because this requires the computationally expensive full system model solution for all parameters in the validation set), the norm of the residual is usually employed as a measure of the discrepancy between the two solutions. Grepl and Patera [21] develop an a posteriori error bound based on the computation of the residual to assess the accuracy of the reduced-model solution. This technique relies on an automatic procedure to establish the basis functions that are required to achieve the reduction at a desired level of accuracy. Pasetto et al. [26] demonstrate that the greedy algorithm is a viable methodology to construct an accurate reduced model for the simulation of groundwater flow in the presence of random transmissivity, when the latter is described by a zonation approach. They show that the number of iterations of the greedy algorithm, which coincides with the number of full system model solutions, is determined by the error tolerance and the number of independent random parameters (i.e., transmissivity zones) considered in the model.

Here, we study on the implementation of the RB approach to construct a reduced-order model of steady-state groundwater flow driven by a randomly distributed transmissivity field characterized by a log-normal probability density. The latter constitutes a multiplicative noise to the flow equation and still represents a critical challenge in modern stochastic hydrogeology. We provide a set of guidelines to establish the appropriate tolerance level to be set in the greedy algorithm by considering a synthetic scenario representing a uniform flow in the mean taking place within a two-dimensional bounded domain. The convergence rate of the greedy algorithm is investigated as a function of the correlation length and variance of the random transmissivity field. We explore the relationship between the number of iterations of the greedy algorithm and the reduced model accuracy upon implementing the algorithm by considering the explicit computation of the error on the basis of a large collection of MC realizations. This allows circumventing inaccuracies related to the use of the residual-based error estimations, which are only partially informative of the head pdf stemming from MC simulations. The reliability of the heuristic criteria underlying the application of the greedy algorithm is assessed in terms of the maximum norm of the error between reduced-order and full system model results. Finally, we assess the accuracy of the reduced-order model by comparing the low-order moments (ensemble mean and variance) and the empirical probability distribution of nodal hydraulic heads resulting from a large set of MC simulations performed with the full system model and the reduced-order model subject to different values of the error tolerance.

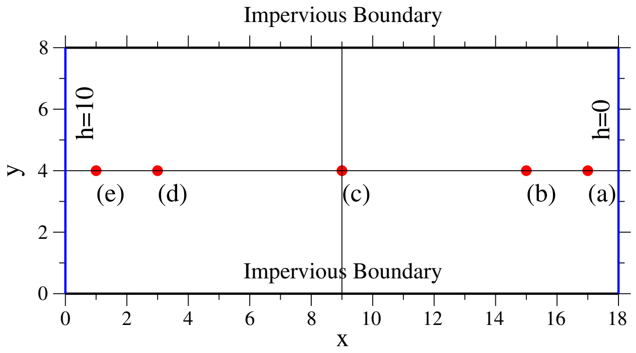

Fig. 1 Two-dimensional domain, S, employed in the numerical simulations. Location of control points considered in Fig. 7 is reported

2 Problem setting

We consider a fully saturated groundwater flow in a porous domain with random hydraulic properties, described by the following stochastic equation:

$$\begin{cases} -\nabla \cdot \left(T\left(\mathbf{x},\omega\right)\nabla h\left(\mathbf{x},\omega\right)\right)=0, & \mathbf{x}\in S \\ h\left(\mathbf{x}\right)=h_{\mathrm{D}}\left(\mathbf{x}\right), & \mathbf{x}\in\partial S_{\mathrm{D}}\subset\partial S \\ -T\left(\mathbf{x},\omega\right)\nabla h\left(\mathbf{x},\omega\right)=q_{\mathrm{N}}\left(\mathbf{x}\right), & \mathbf{x}\in\partial S_{\mathrm{N}}\subset\partial S \end{cases} \qquad (1)$$

where $\mathbf{x}$ is spatial-coordinate vector in the domain S ($S \subset \mathbb{R}^d$, $d = 1, 2$, or 3), ∂S is the boundary of the domain S, h is hydraulic head, ω is a random sample in the space of outcomes Ω, and T is a randomly heterogeneous spatial transmissivity field. The functions h_{D} and q_{N} are the hydraulic heads and Darcy fluxes prescribed at the Dirichlet boundary ∂S_{D} and at the Neumann boundary ∂S_{N}, respectively. We consider T as a stationary stochastic process, characterized by a log-normal distribution, i.e., $Y = \log T$ is a Gaussian random field, with uniform mean μ_Y and covariance function C_Y,

$$C_Y(r) = \sigma_Y^2\, \rho_Y(r), \qquad (2)$$

where r is separation distance (lag), and σ_Y^2 and $\rho_Y(r)$ are the variance and the correlation function of Y, respectively.

The probability space is discretized by means of a number N_{ens} of MC samples. Let $\mathcal{Y}$ be the set of the independent random realizations of Y employed in the MC approach, $\mathcal{Y} = \{Y^{(1)}, \ldots, Y^{(N_{\mathrm{ens}})}\}$ and $\mathcal{T} = \{T^{(1)}, \ldots, T^{(N_{\mathrm{ens}})}\}$ the corresponding transmissivities set. The numerical discretization of the flow problem is obtained by means of the Galerkin finite element method with piecewise linear elements on a triangular grid with n nodes. Solving Eq. 1 for each element $T^{(i)} \in \mathcal{T}$ entails dealing with a high dimensional sparse linear system

$$\mathbf{A}^{(i)}\mathbf{h}^{(i)} = \mathbf{b}, \qquad (3)$$

where $\mathbf{A}^{(i)}$ is the stiffness matrix of dimension $n \times n$ associated with the realization $T^{(i)}$, $\mathbf{h}^{(i)}$ is the vector of hydraulic heads at the grid nodes, and $\mathbf{b}$ is the vector accounting for the boundary conditions. Hereafter, we refer to the solution of Eq. 3 as the full system model (FSM). Given the calculated collection of MC realizations $\{h^{(1)}, \ldots, h^{(N_{\mathrm{ens}})}\}$, an ensemble moment μ of hydraulic head at grid node x_j is approximated by its sample counterpart μ^{FSM}:

$$\mu(h(x_j)) = \int_{\mathbb{R}} \phi\left(h\right)\, p_h\left(h\right)\, dh \approx \mu^{\mathrm{FSM}}\left(h_j\right)$$
$$= \frac{1}{N_{\mathrm{ens}}}\sum_{i=1}^{N_{\mathrm{ens}}} \phi\left(h_j^{(i)}\right), \qquad (4)$$

Fig. 2 Convergence of the maximum relative error $\epsilon_k^{(i)}$ with the number of basis functions N_{BF} employed in the reduced-order model at each iteration of Algorithm 1. Results are illustrated for all 24 scenarios of log-transmissivity considered

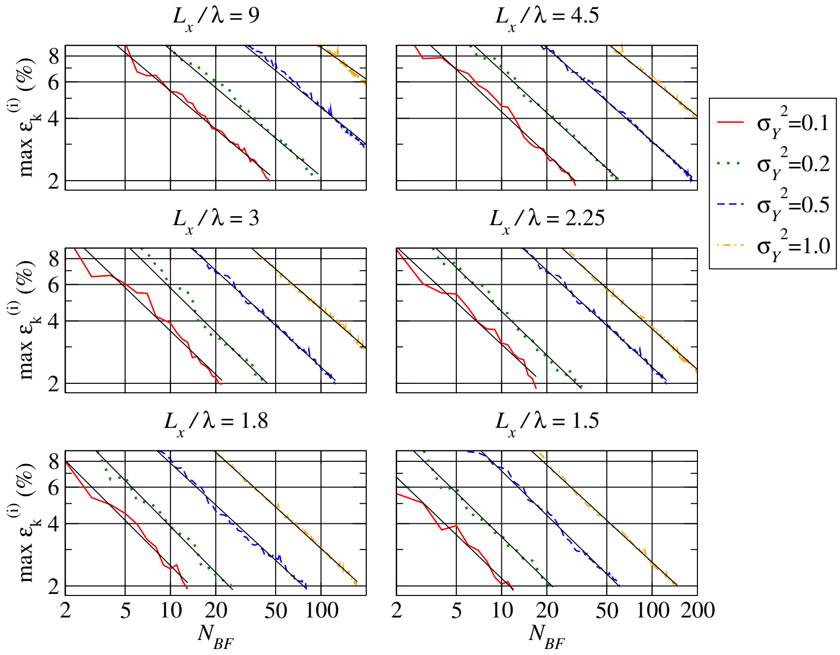

Table 1 Values of the order of convergence m and coefficient c of Eq. 14 for the 24 test cases analyzed

	m				c			
	$\sigma_Y^2 = 0.1$	$\sigma_Y^2 = 0.2$	$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 1.0$	$\sigma_Y^2 = 0.1$	$\sigma_Y^2 = 0.2$	$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 1.0$
$L_x/\lambda = 9$	−0.61	−0.61	−0.59	−0.51	0.223	0.351	0.685	0.935
$L_x/\lambda = 4.5$	−0.67	−0.67	−0.64	−0.58	0.205	0.318	0.609	0.921
$L_x/\lambda = 3$	−0.69	−0.71	−0.66	−0.63	0.177	0.299	0.519	0.854
$L_x/\lambda = 2.25$	−0.67	−0.70	−0.66	−0.65	0.145	0.223	0.519	0.733
$L_x/\lambda = 1.8$	−0.72	−0.73	−0.67	−0.66	0.133	0.211	0.367	0.646
$L_x/\lambda = 1.5$	−0.69	−0.70	−0.71	−0.67	0.108	0.176	0.370	0.574

ϕ and p_h being an integrable function in probability space and the pdf of h at node x_j, respectively.

3 Reduced-order model and greedy algorithm

The reduced-order model is constructed relying on a Galerkin projection technique, which is at the basis of both the RB and POD methodologies. The vector of nodal hydraulic heads is approximated by the sum of a mean head field, $\mathbf{h}^{(0)}$, and a linear combination of N_{BF} spatially distributed basis functions $\mathbf{p}_j$:

$$\mathbf{h}^{(i)} \approx \tilde{\mathbf{h}}^{(i)} = \mathbf{h}^{(0)} + \sum_{j=1}^{N_{\mathrm{BF}}} a_j^{(i)} \mathbf{p}_j = \mathbf{h}^{(0)} + \mathbf{P}\mathbf{a}^{(i)}. \qquad (5)$$

Here, the vector of the coefficients $\mathbf{a}^{(i)} = \{a_1^{(i)}, \ldots, a_{N_{\mathrm{BF}}}^{(i)}\}$ depends on the random realization $T^{(i)}$, and $\mathbf{P}$ is the matrix whose columns are the basis functions $\mathbf{p}_j$. The mean head field $\mathbf{h}^{(0)}$ may be approximated in different ways. For instance, one can employ the solution of Eq. 1 where T is replaced by the geometric mean of the transmissivity field (which provides a relatively robust approximation of the mean head and flux in the absence of forcing terms such as pumping). Alternatively, the solution of the approximated moment equations satisfied by the (ensemble) mean head can be used (see [5]). The coefficients $\mathbf{a}^{(i)}$ are computed in the reduced dimension, i.e., they are the solution of a linear system of dimension $N_{\mathrm{BF}} \times N_{\mathrm{BF}}$ (instead of being the solution of the $n \times n$ FEM system approximating the original groundwater flow model) obtained by projecting Eq. 3 onto the space generated by the N_{BF} basis functions $\mathbf{p}_j$, $j = 1, \ldots, N_{\mathrm{BF}}$. Hence, according to the Galerkin method, we substitute $\mathbf{h}^{(i)}$ with its approximation $\tilde{\mathbf{h}}^{(i)}$ in Eq. 3 and orthogonalize the residual with respect to the basis functions $\mathbf{P}$, obtaining the reduced-order Monte Carlo model (ROMC):

$$\tilde{\mathbf{A}}^{(i)}\mathbf{a}^{(i)} = \tilde{\mathbf{b}}^{(i)}, \qquad (6)$$

where $\tilde{\mathbf{b}}^{(i)} = \mathbf{P}^T\mathbf{b} - \mathbf{P}^T\mathbf{A}^{(i)}\mathbf{h}^{(0)}$, and $\tilde{\mathbf{A}}^{(i)} = \mathbf{P}^T\mathbf{A}^{(i)}\mathbf{P}$ is a symmetric full matrix with dimension $N_{\mathrm{BF}} \times N_{\mathrm{BF}}$.

Solving Eq. 6 is computationally more advantageous than solving the FSM when the number of basis functions is significantly smaller than the number of grid nodes, i.e. $N_{\mathrm{BF}} << n$. This is balanced by the requirement that enough basis functions of sufficient quality should be employed to

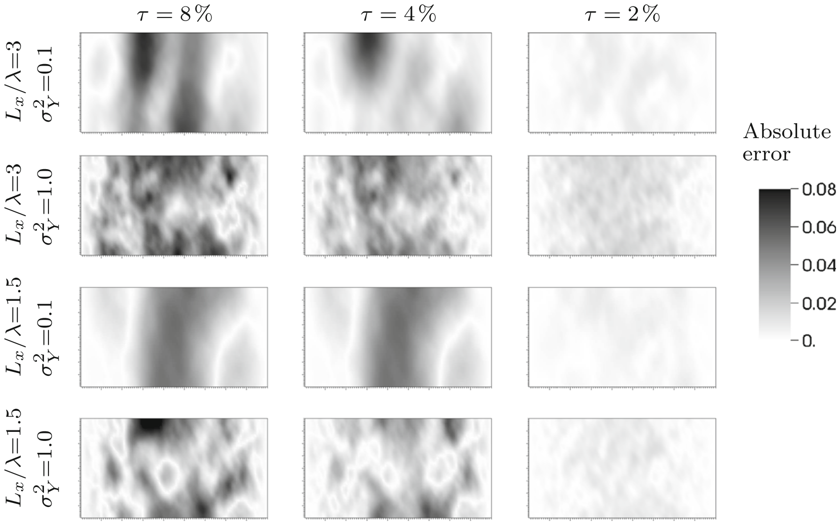

Fig. 3 Spatial distribution of the absolute error between the empirical variances of the hydraulic head realizations computed with the FSM and the ROMC. Results for four test cases (corresponding to all combinations of $L_x/\lambda = 3$, 1.5 and $\sigma_Y^2 = 1.0$, 0.1) and three error tolerances $\tau = 8$ %, 4 %, 2 % are shown

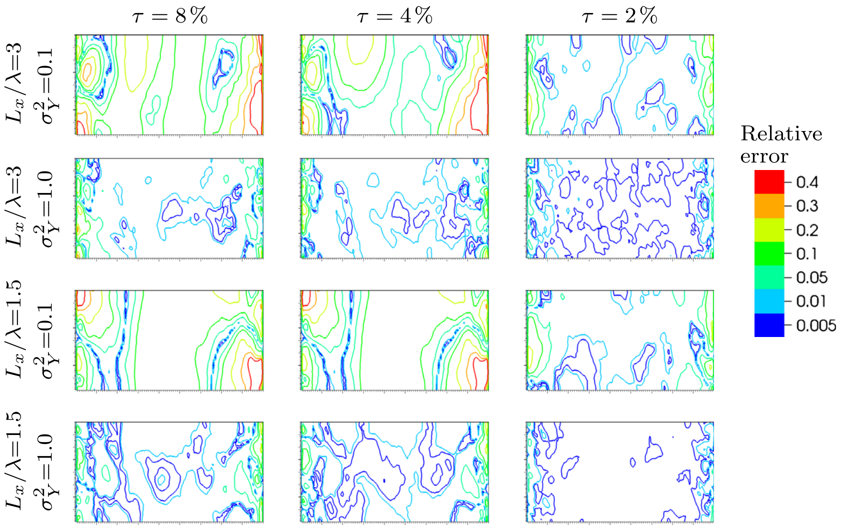

Fig. 4 Spatial distribution of the relative error between the empirical variances of the hydraulic head realizations computed with the FSM and the ROMC. Results for four test cases ($L_x/\lambda = 3$, 1.5 and $\sigma_Y^2 =$ 1.0, 0.1) and three error tolerances $\tau = 8$ %, 4 %, and 2 % are shown

guarantee the accurate reproduction of the empirical pdf evaluated from the MC realizations. Thus, the number of basis functions employed to achieve the reduction and their selection procedure characterize the reduced-order model. Note that the reduction methodology can also be applied jointly with techniques that accelerate the convergence of the MC method, such as sparse grids, sparse approximation, polynomial chaos expansion, latin hypercube, or multilevel MC (see, e.g., [27]).

The POD and RB approaches employ basis functions extracted from a prescribed number of FSM solutions, which are termed snapshots.

In the POD method, the snapshot technique [13] relies on collecting a number N_{snap} of FSM solutions and performing a principal component analysis of these snapshots. Then, the set of basis functions corresponds to the N_{BF} principal components associated with the rightmost singular values and provides the best approximation of the snapshots according to a predefined criterion. The number of snapshots, the criteria for their selection, and the number of principal components to use in the reduction process are (in principle) arbitrary and strongly affect the quality of the reduction.

The RB method employs a different approach: the snapshots are computed using a greedy algorithm and are directly orthonormalized to form the set of basis functions used in the reduction. The greedy algorithm [28] is a deterministic methodology that enables one to obtain a set of suitable snapshots for the RB technique. The main advantage of this procedure with respect the POD method is that it requires only the collection of FSM solutions that are essential for the computation of the basis function set.

The greedy algorithm proceeds iteratively until some suitable metric quantifying the errors between the ROMC and FSM solutions is below a given tolerance τ. At each iteration, the scheme augments the set of the basis functions with the realization associated with the largest error between the full and reduced model solutions. The idea underlying this choice is that the information content embedded in such a realization is not included in the current set of basis functions. Since the computation of the error is costly as it requires the knowledge on the FSM solution, the RB approach employs an a posteriori error estimation to assess the accuracy of the reduced model. If the error estimation is larger then the tolerance τ, the reduced model is enriched with a new basis function; otherwise, the algorithm terminates. The a posteriori error estimation is typically based on the computation of the residual associated with the reduced model solution to preserve the computational efficiency of the algorithm.

Here, we analyze the convergence properties of the greedy algorithm when applied to the MC solution of the stochastic PDE in Eq. 1. As noted in Section 1, a nonaccurate evaluation of the error estimate may lead to an inefficient reduced model. For this reason, our implementation of the greedy algorithm in the RB context relies on the exact computation of the errors instead of considering residual-based error estimates (as described, e.g., in [22]). At each k-th iteration of the greedy algorithm, we compute the ROMC solutions $\left\{\tilde{\mathbf{h}}_k^{(1)}, \dots, \tilde{\mathbf{h}}_k^{(N_{\text{ens}})}\right\}$ using $N_{\text{BF}} = k$ basis functions, $\mathbf{P}_k = [\mathbf{p}_1, \dots, \mathbf{p}_k]$. We consider the relative error measure defined as follows:

$$\epsilon_k^{(i)} = \frac{\|\mathbf{h}^{(i)} - \tilde{\mathbf{h}}_k^{(i)}\|_\infty}{\|\mathbf{h}^{(i)}\|_\infty},$$

i.e., $\epsilon_k^{(i)}$ is the largest nodal error norm calculated between the i-th solution of the FSM and the corresponding ROMC solution normalized by the largest head value of realization i. Let $g_1, \dots, g_k$ be the indices of the head realizations employed as snapshots in the k-th iteration of the greedy algorithm. If the largest value amongst the errors $\left\{\epsilon_k^{(1)}, \dots, \epsilon_k^{(N_{\text{ens}})}\right\}$ is smaller than the predefined tolerance

τ, then the algorithm terminates. Otherwise, the set of basis functions is enriched by employing the FSM solution that is associated with the largest value of the relative error $\epsilon_k^{(i)}$. In other words, at iteration $k+1$, we add the FSM realization characterized by the largest error norm, $\mathbf{h}^{(g_{k+1})}$ to the set of previously calculated snapshots, where

$$g_{k+1} = \underset{1 \le i \le N_{\text{ens}}}{\arg\max}\ \epsilon_k^{(i)}. \tag{7}$$

It follows that $\mathbf{p}_{k+1}$ is the orthonormalization of $\mathbf{h}^{(0)} - \mathbf{h}^{(g_{k+1})}$ with respect to the set of the previously calculated basis functions, $\{\mathbf{p}_1, \dots, \mathbf{p}_k\}$. The initialization of the greedy algorithm is performed by setting $k = 0$, $\tilde{\mathbf{h}}_0^{(i)} = \mathbf{h}^{(0)}$, and

$$g_1 = \underset{1 \le i \le N_{\text{ens}}}{\arg\max}\ \epsilon_0^{(i)}, \qquad \mathbf{p}_1 = \frac{\mathbf{h}^{(0)} - \mathbf{h}^{(g_1)}}{\|\mathbf{h}^{(0)} - \mathbf{h}^{(g_1)}\|_2}. \tag{8}$$

Algorithm 1 illustrates this methodology.

Several techniques are available in the literature to reduce the computational cost associated with the greedy algorithm, as detailed in the following. Since the FSM solutions are typically not available, the norm of the residual term $\mathbf{r}_k^{(i)}$,

$$\mathbf{r}_k^{(i)} = \mathbf{b} - \mathbf{A}^{(i)} \tilde{\mathbf{h}}_k^{(i)} \tag{9}$$

is usually employed for the evaluation of the error $\epsilon_k^{(i)}$ in Eq. 7 (see, e.g., [21, 26]). The evaluation of the residual in

Algorithm 1 Greedy algorithm

```
$k \leftarrow 0$
Compute $\mathbf{h}^{(0)}$
Compute $\left\{\epsilon_0^{(1)}, \dots, \epsilon_0^{(N_{\text{ens}})}\right\}$
while $\max_i \epsilon_k^{(i)} \ge \tau$ do
    $k \leftarrow k+1$
    $g_k = \arg\max_i\ \epsilon_{k-1}^{(i)}$ (Eq. 7)
    $\mathbf{p}_k = \mathbf{h}^{(0)} - \mathbf{h}^{(g_k)}$
    Orthonormalize $\mathbf{p}_k$ with respect to $\{\mathbf{p}_1, \dots, \mathbf{p}_{k-1}\}$
    $\mathbf{P}_k = [\mathbf{p}_1, \dots, \mathbf{p}_k]$
    Compute $\left\{\epsilon_k^{(1)}, \dots, \epsilon_k^{(N_{\text{ens}})}\right\}$
end while
```

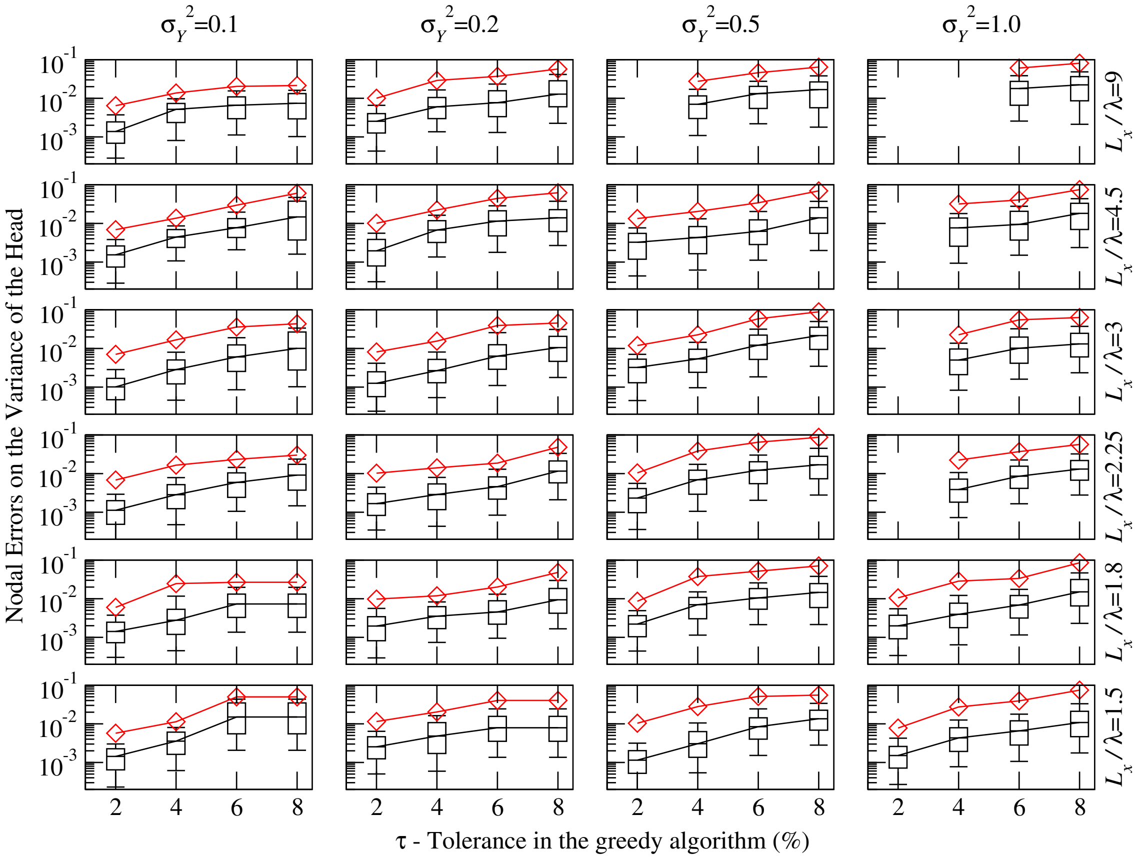

Fig. 5 Distribution of the nodal errors between the empirical hydraulic head variances computed with the FSM and the ROMC. Each *boxplot* depicts the 10th, 25th, 50th, 75th, and 90th percentile of the error distributions. The *red curves* represent the maximum error. Results of test cases that did not attain convergence are omitted

Eq. 9 requires only the computation of the FSM solution of the realizations that maximize the norm of the residual. Note that the norm of the residual can be computed in the reduced-order space when $A^{(i)}$ is expressed as a linear combination of $N_m << n$ matrixes, i.e.,

$$\mathbf{A}^{(i)} = \sum_{j=1}^{N_m} c_j^{(i)} B^{(j)}, \tag{10}$$

where $B^{(1)}, \ldots, B^{(N_m)}$ are parameter-independent matrixes of dimension $n \times n$, while $c_1^{(i)}, \ldots, c_{N_m}^{(i)}$ are coefficients depending on the random realization $T^{(i)}$. This is, e.g., the case when the random field T is approximated through the first N_m terms of the associated Karhunen–Loève expansion. Another procedure to accelerate the greedy algorithm is based on the computation of the basis functions relying either only on a subset of the original MC realizations or on particular samples of the random field [26].

Here, we analyze the way the empirical pdf obtained with the ROMC procedure depends on (a) the tolerance τ adopted in the greedy algorithm, (b) the correlation length, and (c) variance of the log-transmissivity field. Our implementation of the greedy algorithm is based on the evaluation of the error on the FSM, without adding further approximations that can compromise the convergence of the greedy algorithm and the accuracy and efficiency of the resulting ROMC.

Given the approximated realizations $\{\tilde{\mathbf{h}}^{(1)}, \ldots, \tilde{\mathbf{h}}^{(N_{\text{ens}})}\}$ obtained with Algorithm 1 at iteration k (the subscript is omitted for easier reading), the error between a given (ensemble) moment μ^{FSM} (Eq. 4) of the head at node x_j and its counterpart, μ^{ROMC}, calculated on the basis of the sample of ROMC realizations is given by the following:

$$\begin{aligned}&\left|\mu^{\text{FSM}}(h_j) - \mu^{\text{ROMC}}(\tilde{h}_j)\right| \le \textstyle\sum_{i=1}^{N_{\text{ens}}} \frac{\left|\phi\left(h_j^{(i)}\right) - \phi\left(\tilde{h}_j^{(i)}\right)\right|}{N_{\text{ens}}}\\ &= \textstyle\sum_{i=1}^{N_{\text{ens}}} \left|\phi'\left(\xi_j^{(i)}\right)\right| \frac{\left|h_j^{(i)} - \tilde{h}_j^{(i)}\right|}{N_{\text{ens}}} \le C_j \sum_{i=1}^{N_{\text{ens}}} \frac{|\epsilon^{(i)}|}{N_{\text{ens}}} \le C_j\, \tau,\end{aligned} \tag{11}$$

where $\xi_j^{(i)}$ is a point in the interval $\left[h_j^{(i)}, \tilde{h}_j^{(i)}\right]$ that satisfies the mean value theorem and $C_j = \max_i \left|\phi'\left(h_j^{(i)}\right)\right| \; \|\mathbf{h}^{(i)}\|_\infty$. For example, one can note that $C_j = \max_i \|\mathbf{h}^{(i)}\|_\infty$ or $2 \max_i \left|h_j^{(i)}\right| \; \|\mathbf{h}^{(i)}\|_\infty$ for the first and second moments, respectively. Equation 11 guarantees that the sample moments that are computed through the ROMC converge to the FSM moments when τ tends to zero. We also note that the error on the moments (Eq. 11) is bounded by the average nodal error. The latter can be, in turn, significantly smaller than the maximum nodal error and can be used as a termination condition for the greedy algorithm.

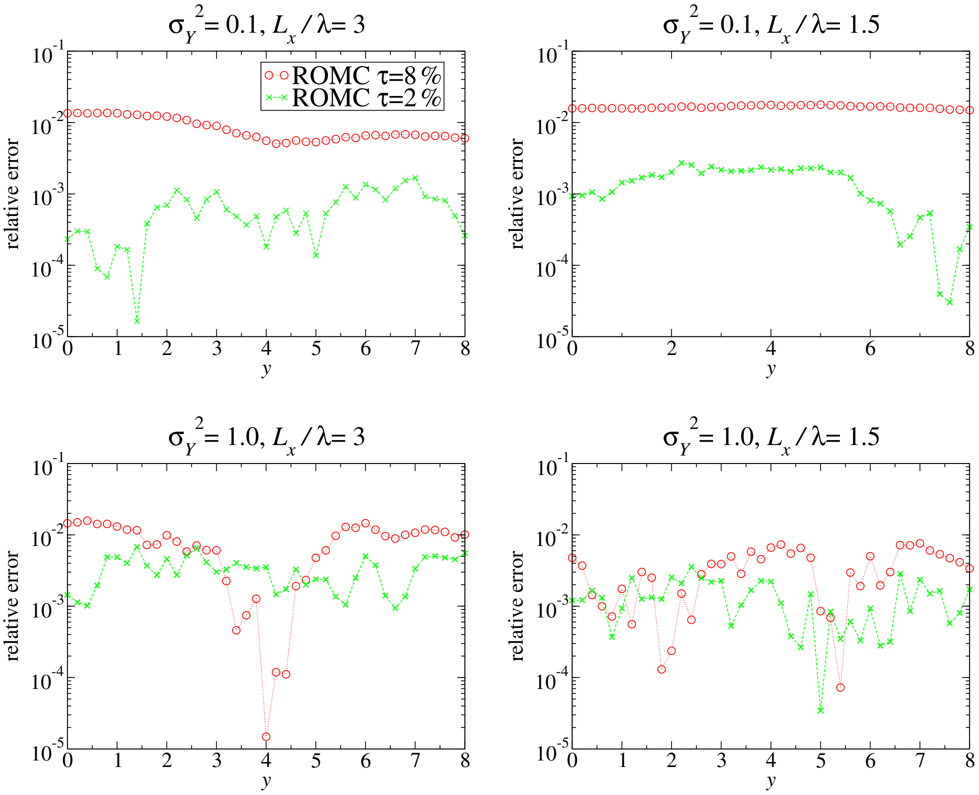

Fig. 6 Relative errors on the 10th percentile, q_{10}, of hydraulic head along the cross-section normal to the mean flow direction and passing through the domain center

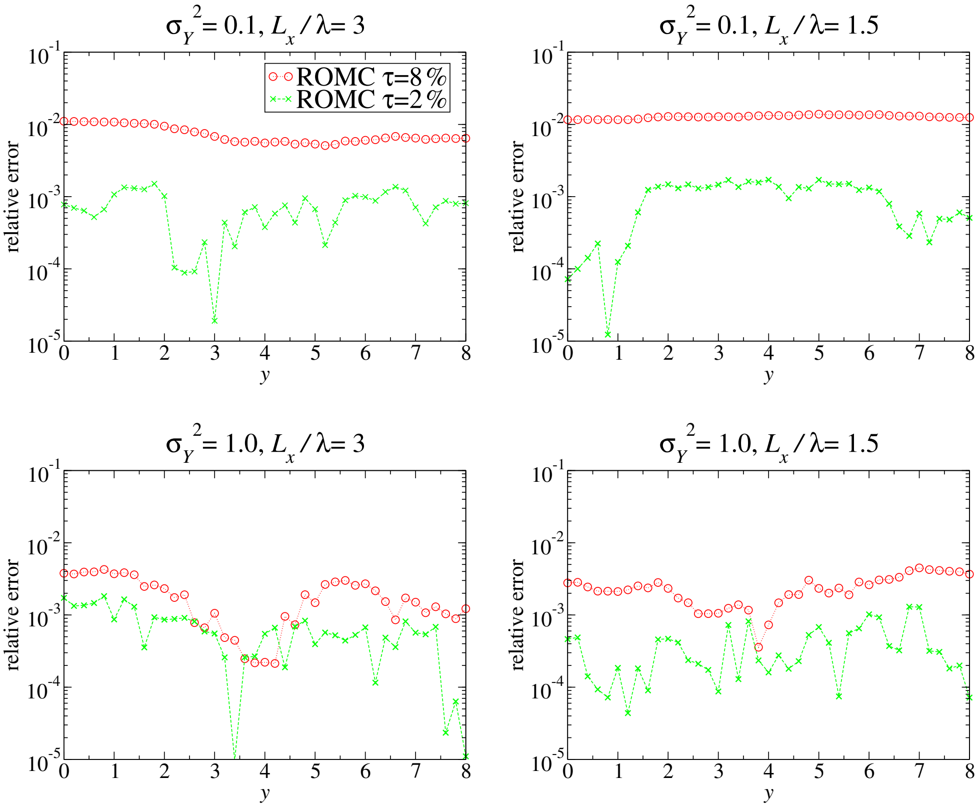

Fig. 7 Relative errors on the 90th percentile, q_{90}, of hydraulic heads along the cross-section normal to the mean flow direction and passing through the domain center

4 Numerical example

We solve Eq. 1 in the two-dimensional domain S depicted in Fig. 1, which has a rectangular shape of length $L_x = 18$ and $L_y = 8$ units in the x and y directions, respectively (here and in the following all quantities are given in consistent units). Prescribed heads of 10 and 0 are imposed at the left and right boundaries, respectively. Impervious boundary conditions are set at the top and bottom boundaries. The domain is discretized into 40 × 90 square cells, for a total of 3,731 nodes.

The stationary log-transmissivity field, Y, is characterized by a normal probability distribution with $\mu_Y = 0$ and exponential isotropic covariance function,

$$C_Y(r) = \sigma_Y^2 \exp\left(-\frac{r}{\lambda}\right), \tag{12}$$

where λ is the correlation length of Y. We analyze 24 test cases characterized by all the combinations of the following values of variance and relative domain size L_x/λ:

$$\sigma_Y^2 = \{0.1, 0.2, 0.5, 1.0\}; \quad \frac{L_x}{\lambda} = \{9, 4.5, 3, 2.25, 1.8, 1.5\}. \tag{13}$$

For each of these cases, we generate 10^4 independent MC realizations of the log-transmissivity field through the sequential Gaussian software HYDRO_GEN [29]. We then solve the FSM (Eq. 3) for each transmissivity field and compute the empirical distribution of the hydraulic head at the grid nodes.

Algorithm 1 is employed for the construction of the reduced-order model. We test four threshold values for the error tolerance, $\tau = \{8\ \%, 6\ \%, 4\ \%, 2\ \%\}$. For each of these tolerances, we compute the collection of ROMC heads and compare their empirical pdf against the one obtained by way of the FSM realizations.

5 Results and discussion

5.1 Convergence of the greedy algorithm

The number of iterations required to attain convergence of the greedy algorithm depends on the error tolerance τ and on the spatial variability of the random realizations of h. Realizations associated with large variability (both in space and in the ensemble sense) require an increased number of basis functions (i.e., of FSM solutions) to obtain approximation errors smaller than the tolerance τ.

Figure 2 depicts (in log–log scale) the convergence of the maximum nodal error $\epsilon_k^{(i)}$ as a function of the number of iterations of the greedy algorithm for the 24 test cases analyzed. For convenience, a maximum number of 200 iterations is set to terminate the algorithm. On one hand, we

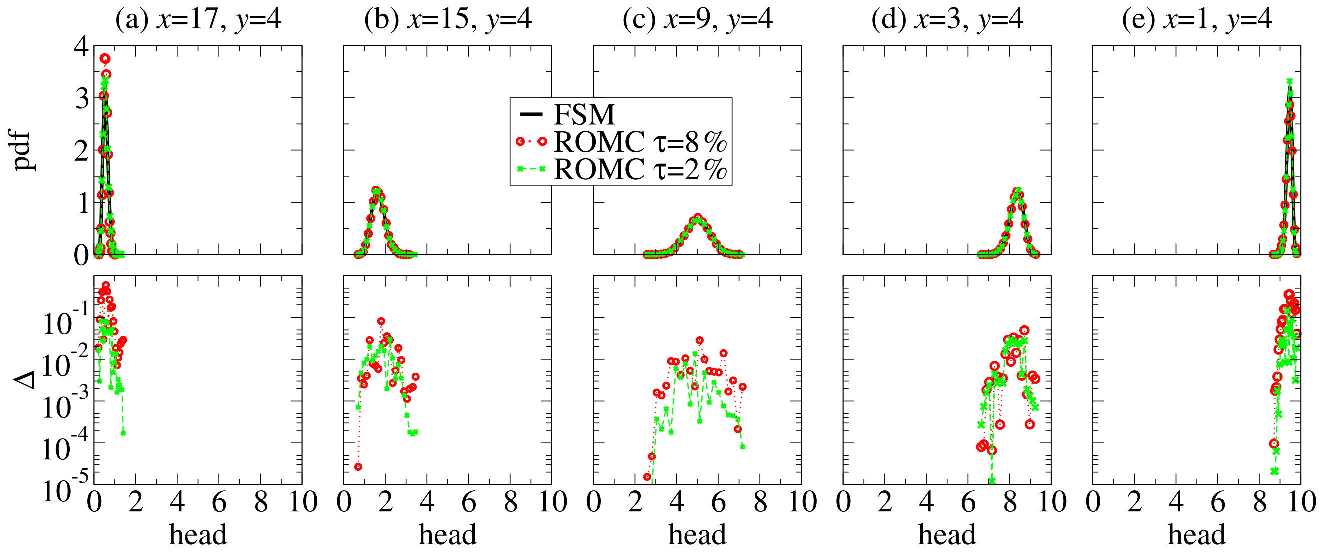

Fig. 8 Comparison between the empirical pdfs of the FSM and ROMC head realizations at the five control points depicted in Fig. 1. Results for $L_x/\lambda = 3$ and $\sigma_Y^2 = 0.1$ and error tolerances $\tau = 8$ %, 2 %. The lower panels show the absolute differences (Δ) between the ROMC and the FSM results

note that after 200 iterations, the maximum error is still larger than the prescribed error tolerance when $\tau = 2$ % for the test cases associated with large variances and large relative domain size (e.g., $L_x/\lambda = 9$, 4.5, 3, 2.25, and $\sigma_Y^2 = 1.0$). On the other hand, convergence is fast for settings with small relative domain size and variance (e.g., $L_x/\lambda = 1.8$, 1.2, and $\sigma_Y^2 = 0.1, 0.2$), and the imposed error tolerance is obtained with less than 30 iterations. Figure 2 also reports the regression lines that approximate the convergence rate of the algorithm. These results reveal that the convergence rate of the greedy algorithm is well approximated by a power-law model of the kind:

$$\max_i \left(\epsilon_k^{(i)}\right) \approx c\, k^m. \tag{14}$$

Table 1 summarizes the estimated values of c and m for the 24 test cases. The order of convergence m slightly oscillates about a mean value of −0.66 and lies within the range [−0.73, −0.51]. The coefficient c displays an approximately linear dependence on the relative domain size L_x/λ and variance, its values increasing from 0.108 in the case with $L_x/\lambda = 1.5$, $\sigma_Y^2 = 0.1$ to 0.935 for $L_x/\lambda = 9$, $\sigma_Y^2 = 1.0$.

These results may be explained by the observation that large correlation lengths of the Y field (i.e., low values of L_x/λ) are associated with relatively smooth spatial variations of h. This implies that the ensemble variance of hydraulic heads tends to decrease with decreasing L_x/λ for the flow condition considered. As a consequence, only few basis functions are required to describe the realizations of head, and the greedy algorithm converges more rapidly. A corresponding argument explains the behavior observed for the settings associated with low variance of Y.

5.2 Estimation of head variance

Here, we analyze the accuracy of the ROMC procedure for the approximation of the empirical variance of hydraulic heads. Figures 3 and 4, respectively, depict the spatial distribution of the absolute and relative errors between the empirical variances computed via the FSM and the ROMC. Results are illustrated for four selected test cases (corresponding to all combinations of $L_x/\lambda = 3$, 1.5, and $\sigma_Y^2 = 1.0$, 0.1) and three error tolerances $\tau = 8$ %, 4 %, and 2 %. In agreement with Eq. 11, the error on the variance decreases throughout the domain as τ decreases, its values always being of the same order of magnitude as those of τ (see Fig. 3). This in turn implies that Eq. 11 can be considered as a conservative upper bound on the error performed by the ROMC on the computation of the second moment.

Note that the magnitude of the absolute errors depicted in Fig. 3 does not display large variations for a given error tolerance. This implies that τ controls the accuracy of the reproduction of the collection of MC realizations and their statistics. This result is consistent with Eq. 11, where it is seen that the error is independent of the correlation length and variance of Y. The generality of this conclusion within the range of parameters tested is supported by Fig. 5, where the boxplots of the distribution of the spatial nodal errors on the head variance are shown for the complete set of test cases and the four error tolerances analyzed. Figure 4 reveals that the largest relative nodal errors occur in the proximity of the Dirichlet boundary, i.e., where the head

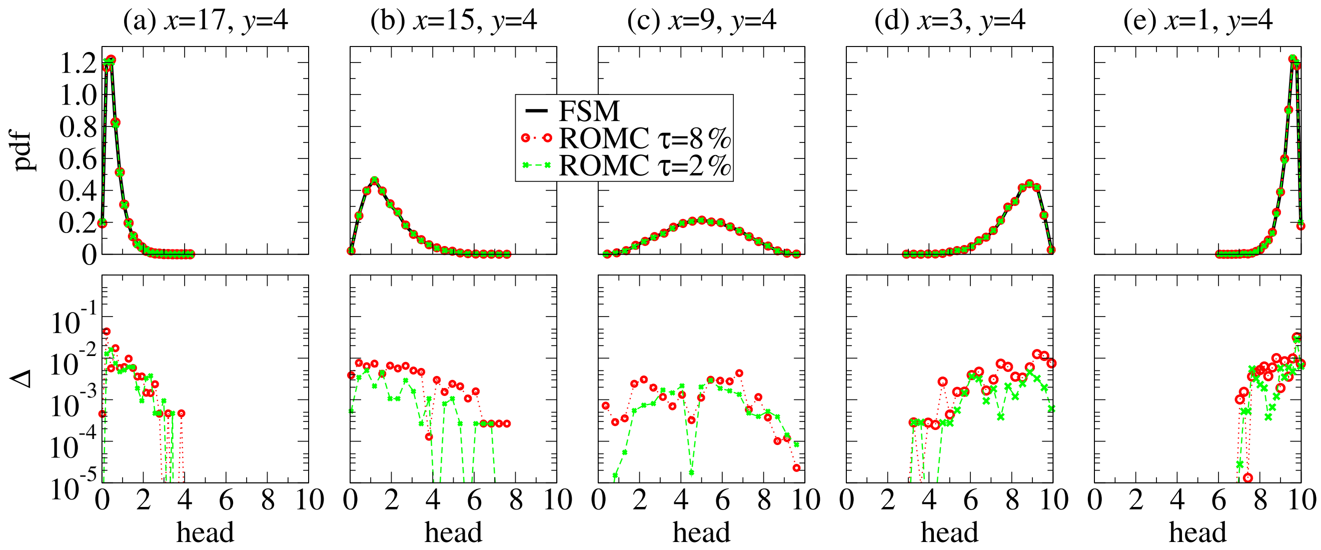

Fig. 9 Comparison between the complete empirical pdfs of the FSM and ROMC head realizations at the five control points depicted in Fig. 1. Results for $L_x/\lambda = 3$ and $\sigma_Y^2 = 1.0$ and error tolerances $\tau = 8$ %, 2 %. The *lower panels* show the absolute differences (Δ) between the ROMC and the FSM results

variance tends to vanish, while they are always smaller than the desired model accuracy at all other locations in the domain.

5.3 Empirical distributions of nodal heads

MC simulations are frequently employed (e.g., in environmental risk assessment protocols) to explore the behavior at the tails of the probability distribution. Here, we focus on the ability of the ROMC scheme to approximate the 10th and 90th percentiles of nodal heads distributions, q_{10} and q_{90}, respectively, defined as follows:

$$P\,(h < q_{10}) = 10\ \% \ , \qquad P\,(h < q_{90}) = 90\ \% \qquad (15)$$

where P indicates probability. Figures 6 and 7 report the relative errors on the q_{10} and q_{90} for hydraulic heads at nodes located along the central vertical cross-section of the domain, where head variances are highest. By way of illustration and consistent with Figs. 3 and 4, the results are presented for the test cases associated with all combinations of $L_x/\lambda = 3$, 1.5 and $\sigma_Y^2 = 1.0,\ 0.1$. The comparison of the calculated percentiles associated with the error tolerances $\tau = 8$ % and $\tau = 2$ % shows that the smallest error tolerance renders the best reproduction of the FSM percentiles. The maximum relative errors in both the 10th and 90th percentiles are associated with $\sigma_Y^2 = 0.1$ and $\tau = 8$ %. This is likely related to the small number of basis functions returned by the greedy algorithm in these two cases ($N_{\mathrm{BF}} = 3$ for $L_x/\lambda = 3$ and $N_{\mathrm{BF}} = 2$ for $L_x/\lambda = 1.5$). We note that this effect is not visible for $\tau = 2$ %, where the errors on the percentiles are of the same order of magnitude for all tested cases. Similarly to what was observed for the errors on the head variance (Section 5.3), the magnitude of the relative errors on the percentiles are significantly smaller than the error tolerances τ at most locations in the domain.

Finally, Figs. 8 and 9 present a comparison between the complete empirical pdfs of the FSM and ROMC head realizations at five selected points in the system. Results are illustrated for two test cases ($L_x/\lambda = 3$, $\sigma_Y^2 = 0.1$ or 1.0) and error tolerances $\tau = 8$ %, 2 %. The small differences between the pdfs, which have also been detected in the analysis of the head variance and percentiles, do not appear to be significant when we consider the complete range of the hydraulic heads [$0 \leq h \leq 10$] resulting from the collection of MC realizations.

Results of comparable quality are obtained also for the remaining test cases (not shown).

6 Conclusions

Our work leads to the following major conclusions.

- The greedy algorithm is a suitable procedure to construct a reduced-order model for the solution of steady-state groundwater flow in the presence of randomly distributed (Gaussian) log-transmissivities characterized by relatively low variance and highly persistent spatial distribution (i.e., large correlation length scales relative to the size of the domain). Our examples show that under these conditions the algorithm reduces the original FSM of dimension $n = 3,371$ to a RM of dimension $N_{\mathrm{BF}} < 50$ maintaining a high level of accuracy on each realization of hydraulic heads (see Fig. 2, for the test cases with $\sigma_Y^2 < 0.5$ and $L_x/\lambda < 3$, with $\tau = 2\ \%$). In these cases, the procedure requires only a few FSM runs and is markedly advantageous with respect to the computational cost associated with the classical MC scheme (10^4 FSM runs in our case). An efficient and accurate reduction for the configurations characterized by higher variances and smaller correlation lengths requires setting increased values of the error tolerance τ, even though this compromises the accuracy on the reproduction of the head realizations and the empirical head distributions (Figs. 3, 4, and 5).

- The rate of convergence of the largest error on the head realizations can be approximated as a power law function of the number of iterations of the greedy algorithm (Fig. 2, Table 1). This relationship between errors and number of iterations can be considered as an "a priori" criterion to establish the number of iterations of the greedy algorithm that are required to reach the desired error tolerance. Note that, since the implementation of the greedy algorithm with the residual-based estimation of the error may lead to a selection of snapshots that do not minimize the discrepancy between the FSM and ROMC, the convergence rate of the algorithm that employs residual-based error estimations may be slower than the one presented here.
- The collection of hydraulic head realizations computed with the reduced-order model is associated with an empirical probability distribution that well approximates the FSM-based sample distribution. We demonstrate this by showing comparisons of the spatial variance of heads (Figs. 3, 4, and 5), the spatial distribution of head percentiles along selected domain cross sections (Figs. 6 and 7) and the entire empirical pdf at selected grid nodes (Figs. 8 and 9) calculated through the FSM and RM set of realizations. As expected, the error on the ensemble statistics decreases with decreasing error tolerance in the greedy algorithm. It can be noted that, once the error tolerance τ is assigned, the accuracy of the estimation of the head statistics is not significantly deteriorated by changes in the geostatistical parameters that describe the random spatial distribution of the log-transmissivity.

Acknowledgments Funding from MIUR (Italian Ministry of Education, Universities and Research—PRIN2010-11; project "Innovative methods for water resources under hydro-climatic uncertainty scenarios"), from EU FP7 (project CLIMB "Climate Induced Changes on the Hydrology of Mediterranean Basins—Reducing Uncertainty and Quantifying Risk"), and from the Fondazione CaRiPaRO (project "Nonlinear Partial Differential Equations: models, analysis, and control-theoretic problems") is acknowledged.

References

1. Dagan, G.: Flow and Transport in Porous Formations. Springer, New York (1989)
2. Zhang, D.: Stochastic Methods for Flow in Porous Media: Copying with Uncertainties. Academic, San Diego (2002)
3. Yustres, A., Asensio, L., Alonso, J., Navarro, V.: A review of Markov Chain Monte Carlo and information theory tools for inverse problems in subsurface flow. Comput. Geosci. **16**(1), 1–20 (2012). doi:10.1007/s10596-011-9249-z
4. Moradkhani, H., DeChant, C.M., Sorooshian, S.: Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. Water Resour. Res. **48**(12), W12520 (2012). doi:10.1029/2012WR012144
5. Guadagnini, A., Neuman, S.P.: Nonlocal and localized analyses of conditional mean steady state flow in bounded, randomly nonuniform domains: 1. Theory and computational approach. Water Resour. Res. **35**(10), 2999–3018 (1999). doi:10.1029/1999WR900160
6. Winter, C.L., Tartakovsky, D.M., Guadagnini, A.: Moment differential equations for flow in highly heterogeneous porous media. Surv. Geophys. **24**(1), 81–106 (2003). doi:10.1023/A:1022277418570
7. Panzeri, M., Riva, M., Guadagnini, A., Neuman, S.P.: Data assimilation and parameter estimation via ensemble Kalman filter coupled with stochastic moment equations of transient groundwater flow. Water Resour. Res. **49**, 1334–1344 (2013). doi:10.1002/wrcr.20113
8. Sanchez-Vila, X., Fernandez-Garcia, D., Guadagnini, A.: Conditional probability density functions of concentrations for mixing-controlled reactive transport in heterogeneous aquifers. Math. Geosci. **41**, 323–351 (2009). doi:10.1007/s11004-008-9204-2
9. Tartakovsky, D.M., Dentz, M., Lichtner, P.C.: Probability density functions for advective-reactive transport in porous media with uncertain reaction rates. Water Resour. Res. **45**, W07414 (2009). doi:10.1029/2008WR007383
10. Dentz, M., Tartakovsky, D.M.: Probability density functions for passive scalars dispersed in random velocity fields. Geophys. Res. Lett. **45**, L24406 (2010). doi:10.1029/2010GL045748
11. Venturi, D., Tartakovsky, D.M., Tartakovsky, A.M., Karniadakis, G.E.: Exact pdf equations and closure approximations for advective-reactive transport. J. Comput. Phys. **243**, 323–343 (2013). doi:10.1016/j.jcp.2013.03.001
12. Ballio, F., Guadagnini, A.: Convergence assessment of numerical Monte Carlo simulations in groundwater hydrology. Water Resour. Res. **40**, W04603 (2004). doi:10.1029/2003WR002876
13. Pasetto, D., Guadagnini, A., Putti, M.: POD-based Monte Carlo approach for the solution of regional scale groundwater flow driven by randomly distributed recharge. Adv. Water Resour. **34**(11), 1450–1463 (2011). doi:10.1016/j.advwatres.2011.07.003
14. Zhang, D., Lu, Z.: An efficient, high-order perturbation approach for flow in random porous media via Karhunen-Loève and polynomial expansions. J. Comput. Phys. **194**(2), 773–794 (2004). doi:10.1016/j.jcp.2003.09.015

15. Poles, S., Lovison, A.: A polynomial chaos approach to robust multiobjective optimization. In: Deb, K., et al. (eds.) Hybrid and Robust Approaches to Multiobjective Optimization, number 09041 in Dagstuhl Seminar Proceedings. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, Dagstuhl (2009)
16. Oladyshkin, S., Class, H., Helmig, R., Nowak, W.: An integrative approach to robust design and probabilistic risk assessment for CO2 storage in geological formations. Comput. Geosci. **15**(3), 565–577 (2011). doi:10.1007/s10596-011-9224-8
17. Formaggia, L., Guadagnini, A., Imperiali, I., Lever, V., Porta, G., Riva, M., Scotti, A., Tamellini, L.: Global sensitivity analysis through polynomial chaos expansion of a basin-scale geochemical compaction model. Comput. Geosci. **17**, 25–42 (2013). doi:10.1007/s10596-012-9311-5
18. Li, H., Zhang, D.: Probabilistic collocation method for flow in porous media: comparisons with other stochastic methods. Water Resour. Res. **43**, W09409 (2007). doi:10.1029/2006WR005673
19. Kunisch, K., Volkwein, S.: Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. SIAM J. Numer. Anal. **40**, 492–515 (2002). doi:10.1137/S0036142900382612
20. Siade, A.J., Putti, M., Yeh, W.W.-G.: Snapshot selection for groundwater model reduction using proper orthogonal decomposition. Water Resour. Res. **46**, W08539 (2010). doi:10.1029/2009WR008792
21. Grepl, M.A., Patera, A.T.: A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations. ESAIM-Math Model. Num. **39**(1), 157–181 (2005). doi:10.1051/m2an:2005006
22. Quarteroni, A., Rozza, G., Manzoni, A.: Certified reduced basis approximation for parametrized partial differential equations and applications. Math. Indust. **1**(1), 1–49 (2011). doi:10.1186/2190-5983-1-3
23. Siade, A.J., Putti, M., Yeh, W.W.-G.: Reduced order parameter estimation using quasilinearization and quadratic programming. Water Resour. Res. **48**, W06502 (2012). doi:10.1029/2011WR011471
24. Kaleta, M.P., Hanea, R.G., Heemink, A.W., Jansen, J.-D.: Model-reduced gradient-based history matching. Comput. Geosci. **15**(1), 135–153 (2011). doi:10.1007/s10596-010-9203-5
25. van Doren, J.F.M., Markovinović, R., Jansen, J.-D.: Reduced-order optimal control of water flooding using proper orthogonal decomposition. Comput. Geosci. **10**(1), 137–158 (2006). doi:10.1007/s10596-005-9014-2
26. Pasetto, D., Putti, M., Yeh, W.W.-G.: A reduced order model for groundwater flow equation with random hydraulic conductivity: application to Monte Carlo methods. Water Resour. Res. **49**, 1–14 (2013). doi:10.1002/wrcr.20136
27. Müller, F., Jenny, P., Meyer, D.W.: Multilevel Monte Carlo for two phase flow and Buckley-Leverett transport in random heterogeneous porous media. J. Comput. Phys. (2013). doi:10.1016/j.jcp.2013.03.023
28. Lieberman, C., Willcox, K., Ghattas, O.: Parameter and state model reduction for large-scale statistical inverse problems. SIAM J. Sci. Comput. **32**(5), 2523–2542 (2010). doi:10.2307/2236101
29. Bellin, A., Rubin, Y.: HYDRO_GEN: a spatially distributed random field generator for correlated properties. Stoch. Hydrol. Hydraul. **10**(4), 253–278 (1996). doi:10.1007/BF01581869