https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Joint characterization of heterogeneous conductivity fields and pumping well attributes through iterative ensemble smoother with a reduced-order modeling strategy for solute transport

Chuan-An Xia[1], Jiayun Li[2*], Bill X. Hu[3], Alberto Guadagnini[4,5], Monica Riva[4*]

[1]Zijin School of Geology and Mining, Fuzhou University, Fuzhou, China

[2]Fujian Provincial Key Lab of Coastal Basin Environment, Fujian Polytechnic Normal University, Fuqing, China

[3]School of Water Conservancy & Environment, University of Jinan, Jinan, China

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

[5]Sonny Astani Department of Civil and Environmental Engineering, Viterbi School of Engineering, Los Angeles, California 90089-2531, USA

Corresponding author: Jiayun Li; Monica Riva

Email: lijy@fpnu.edu.cn; monica.riva@polimi.it

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Abstract

We develop and test an efficient and accurate theoretical and computational framework to jointly estimate spatially variable hydraulic conductivity and identify unknown pumping well locations and rates in a two-dimensional confined aquifer. The approach (denoted as iES_ROM) integrates an iterative Ensemble Smoother (iES) with a Reduced-Order Model (ROM) for solute transport taking place across an otherwise steady-state groundwater flow field. This offers a computationally efficient alternative to the Full System Model (iES_FSM) upon addressing the high computational demands of ensemble-based data assimilation methods, which typically require large ensemble sizes to characterize uncertainties in (randomly) heterogeneous aquifers. Our iES_ROM is constructed through proper orthogonal decomposition. It is then evaluated across a collection of 28 test cases exploring variations in model dimension, ensemble size, measurement noise, monitoring network, and statistical properties of the (underlying randomly heterogeneous) conductivity field. Our results support the ability of iES_ROM to accurately estimate conductivity and identify pumping well attributes under diverse configurations, attaining a quality of performance similar to iES_FSM. When using moderate ROM dimensions (n = 25-30) and ensemble size (i.e., 500-1000), the accuracy of iES_ROM does not vary significantly while computational time is reduced by nearly an order of magnitude. Our approach thus provides a reliable and cost-effective tool for inverse modeling in groundwater systems with uncertain parameters.

Keywords: reduced-order model; proper orthogonal decomposition; iterative

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

ensemble smoother; pumping well identification; groundwater

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

1. Introduction

Assessment of groundwater flow and transport scenarios is typically plagued by uncertainties associated with model structure and parametrization. A major source of uncertainty often examined concerns the poorly constrained assessment of pollution sources. Our ability to identify spatial locations of these sources exerts significant influence on the design of contaminant monitoring, management, and remediation strategies. Contaminant release to an aquifer is characterized through the spatial location of sources, the temporal variability of release fluxes, and solute concentrations involved (Chen et al., 2018; Xu et al., 2018; Mo et al., 2019). Uncertainties linked to groundwater abstraction scheduling also play a critical role, as operational details of pumping wells are not always fully documented. For example, this might correspond to a scenario where such information is not disclosed to ensure privacy protection or uncertainties are induced by geocoding practices and/or measurement devices. Further to this, in some regions groundwater may be accessed through wells that are not officially registered or fully documented by industrial operators and/or local residents. Despite the relevance of these issues, only limited research has been devoted to the identification and quantification of pumping rates and spatial locations of such hidden wells.

In this broad context, we recall that a considerable body of research has focused on estimating key parameters (such as hydraulic conductivity) in groundwater flow and transport models through ensemble-based Data Assimilation (DA) techniques (Chen and Zhang, 2006; Tong et al., 2013; Chen and Oliver, 2013; Zhang et al., 2018;

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Xia et al., 2018, 2024). These approaches aim at enhancing the accuracy of simulated system states (e.g., hydraulic heads and solute concentrations). While their capability to jointly estimate parameters and update system states has been broadly explored, their high computational cost still constitutes a persistent limitation to their practical routine use. This challenge primarily stems from the requirement for a large number of realizations to ensure statistical convergence of Monte Carlo (MC) simulations (e.g., Ballio and Guadagnini, 2004) in the forecast step of the DA process, and to achieve reliable parameter estimates in the analysis step. The computational burden becomes particularly significant when the selected model describing the system behavior (hereafter termed as Full System Model (FSM)) must be repeatedly executed for systems characterized by strong nonlinearities or requiring high (space-time) resolution of state variables and parameters.

To alleviate these computational constraints, recent studies explore the benefit of relying on surrogate (or reduced-order) models that approximate the behavior of the full system while maintaining sufficient accuracy for inverse modeling workflows and Uncertainty Quantification (UQ).

In this framework, efforts to mitigate computational limitations of (ensemble-based) DA methods primarily focus on the adoption of localization techniques (e.g., Xia et al., 2018, 2024; Luo and Bahkta, 2020) or surrogate modeling strategies (e.g., Zhang et al., 2018; Mo et al., 2019 and references therein). Main advantages associated with location approaches are related to the observation that they (*i*) substantially reduce computational costs upon requiring only a limited

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

number of Monte Carlo realizations of the FSM, while maintaining acceptable accuracy of the assimilated results, and (*ii*) retain a physically-based and mathematically-tractable formulation. As a notable drawback of these approaches, we note that the value of the information associated with diverse measurements may be partially suppressed due to the use of distance- or correlation-based localization, which might constrain the strength of the spatial influence of observations. As a consequence, the ensuing (empirical/sample) probability density functions (PDFs) of model parameters and system states often display reduced accuracy and fail to fully capture the underlying uncertainty structure. To mitigate these limitations, an alternative line of research explores the use of surrogate models (SMs), which aim at emulating the response of the Full System Model with significantly reduced computational cost while preserving the salient physics of the system.

Surrogate models are rapidly emerging as a promising complement to FSMs for reducing computational burdens associated with the forecast steps of ensemble-based DA procedures. Among the various SM strategies, data-driven approaches based on machine learning (e.g., Ju et al., 2018) and deep learning (e.g., Mo et al., 2019) can be employed for emulating groundwater flow and transport processes taking place in heterogeneous media. For example, Ju et al. (2018) rely on Gaussian Process regression to describe relationships between the coefficients of a Karhunen-Loève (KL) expansion (employed to characterize a spatially heterogeneous hydraulic conductivity field) and (point-wise) simulated observations. This approach is shown to achieve approximately an order of magnitude reduction in computational time as

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

compared with the standard iterative Ensemble Smoother (iES). Otherwise, this gain in efficiency is associated with a reduced accuracy in simulated hydraulic heads, which in turn compromises the reliability of the estimated conductivity field. Mo et al. (2019) employ deep autoregressive neural networks as an FSM surrogate to reconstruct conductivity fields and identify contaminant source characteristics. However, their approach still requires a significant computational effort, as it heavily relies on a high number (about 1,500 in their exemplary setting) of MC realizations of the FSM for network training. While these studies show a clear potential of data-driven surrogates for accelerating DA workflows, they also highlight the need for a fundamental trade-off between computational efficiency and model accuracy, thus underscoring the potential value of alternative surrogate modeling strategies.

In contrast to data-driven models, that typically operate as black-box representations, projection-based Reduced-Order Models (ROMs) are physics-based (e.g., Razavi et al., 2012; Asher et al., 2015; Chen et al., 2017; Xia et al., 2020, 2025). ROMs are typically constructed upon projecting the governing equations and boundary conditions of the onto a lower-dimensional subspace spanned by a set of basis functions. The latter are commonly derived through, e.g., Proper Orthogonal Decomposition (POD) of multiple FSM solutions, referred to as *snapshots*. This procedure effectively reduces the dimensionality of the system state space. The random field representing the system state can then be expressed as a linear combination of the dominant eigenfunctions obtained from the Fredholm integral equation associated with the covariance matrix of the snapshots. Leading

© Author(s) 2025. CC BY 4.0 License.

eigenfunctions are then identified as the basis functions defining the reduced subspace. Substantial computational savings are then achieved upon resting on the solution of the ensuing low-dimensional linear system. When implemented in the context of numerical MC frameworks, the collection of ROM-generated solutions constitutes what is commonly referred to as a Reduced-Order Monte Carlo (ROMC) simulation framework.

Reduced-order modeling has received growing attention in the context of groundwater flow (Pasetto et al., 2011, 2013, 2014; Li et al., 2013a; Boyce et al., 2015; Stanko et al., 2016; Xia et al., 2020, 2025) and solute transport (Luo et al., 2012; Li et al., 2013b; Rizzo et al., 2018) scenarios. Its potential is evidenced across a wide range of hydrogeological configurations, including confined (e.g., Pasetto et al., 2011) and unconfined (e.g., Stanko et al., 2016) aquifer systems, homogeneous (e.g., Li et al., 2013a) and heterogeneous (e.g., Pasetto et al., 2013) media, as well as scenarios with (e.g., Xia et al., 2020) or without (e.g., Pasetto et al., 2014) pumping wells operating therein. Several studies further advance development of ROMC strategies for UQ in groundwater flow modeling. Pasetto et al. (2014) show that the accuracy of UQ results relying on ROMC in the presence of steady-state groundwater flow strongly depends on the quality and the number of snapshots, the latter directly influencing representativeness of the basis functions. To mitigate this limitation, Xia et al. (2020) propose deriving basis functions as the leading eigenvectors of (second-order) approximations of hydraulic head covariances. The latter are obtained upon solving the associated moment equations for steady-state groundwater flow (Zhang and Lu,

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

2002; Xia et al., 2019). Even as reduction of the dimensionality of the head space provides substantial computational savings, projection of the basis functions onto the ensuing (typically large) system matrix remains computationally intensive, thereby still constituting a limiting factor to efficiency gains. Xia et al. (2025) address this challenge by extending their approach to perform dimensionality reduction for both (spatially variable) transmissivity and hydraulic head fields in a steady-state groundwater flow setting and achieving additional computational savings while maintaining high accuracy. Despite these advancements, most existing ROM and ROMC approaches are still fraught with difficulties in efficiently capturing strongly nonlinear system dynamics and adapting to evolving state conditions, underscoring the need for more flexible and computationally efficient reduced-order frameworks.

With reference to solute transport, ROMs have been developed for both homogeneous (Luo et al., 2012) and heterogeneous (Li et al., 2013b; Rizzo et al., 2018) aquifer systems. Li et al. (2013a) further consider construction of ROMs to tackle density-dependent groundwater flow taking place across homogeneous and heterogeneous domains. Otherwise, studies explicitly focusing on the development of ROMC approaches for UQ of solute transport remain limited. Although conceptual insights can be drawn from ROMC studies addressing groundwater flow (e.g., Pasetto et al., 2014; Xia et al., 2020, 2025), influence of key factors (such as, e.g., dimensionality of the reduced concentration space and strength of hydraulic conductivity heterogeneity) on accuracy and robustness of ROMC-based UQ still remains poorly characterized.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Building upon these works, the present study introduces a novel framework that integrates the iES with a ROM for solute transport (hereafter referred to as iES_ROM). The ensuing framework enables one to efficiently quantify uncertainty and jointly estimate system parameters in groundwater-related modeling scenarios. The proposed method is then applied to simultaneously identify pumping rate and spatial location of (otherwise hidden) wells operating within the system, while providing estimates of the spatially heterogeneous hydraulic conductivity field under conditions of steady-state flow and transient solute transport. In the iES_ROM framework, the steady-state flow field is evaluated through the FSM, whereas the transient solute transport is represented by a computationally efficient ROM. The required snapshots and associated POD are generated only once. These are subsequently employed throughout the entire DA process, thus avoiding repeated high-fidelity simulations. To ensure transparent benchmarking, the performance of iES_ROM is systematically compared with that of a reference approach (termed iES_FSM) which relies entirely on the FSM associated with synthetic scenarios. Comparative analyses are performed across a variety of synthetic scenarios, encompassing diverse ROM dimensions, ensemble sizes, measurement qualities and quantities, as well as distinct statistical descriptors of the initial conductivity ensemble and snapshot sizes.

The study is organized as follows. Section 2 introduces the theoretical background of groundwater flow and solute transport and details the integration of ROMC simulation within the iES framework. Section 3 describes the test cases

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

designed to evaluate the proposed approach. Section 4 illustrates and discusses the main results, and Section 5 summarizes the key findings.

2. Theory background and methodology

2.1 Groundwater flow and solute transport

We consider two-dimensional steady-state groundwater flow governed by:

$$\nabla \cdot \left[K(\mathbf{x}) \nabla h(\mathbf{x}) \right] + q_s(\mathbf{x}) = 0 \tag{1}$$

where $\mathbf{x} = [x_1, x_2]$ is a vector of spatial coordinates in domain Ω^2; h is hydraulic head; K is (isotropic) hydraulic conductivity; and q_s is a source/sink term. We conceptualize K as a spatially heterogeneous random field, associated with a given spatial correlation structure. The source/sink term in Equation (1) corresponds to a production well associated with an uncertain pumping rate and location in the domain. Propagation of uncertainty related to model parameters and/or forcing terms onto hydraulic heads and fluxes is typically assessed through numerical Monte Carlo (MC) simulations (see, e.g., Ballio and Guadagnini, 2004; Xia et al., 2020, 2024, and references therein).

We consider (non-reactive) solute transport evolving in Ω^2 to be described through:

$$\nabla \cdot \left[D \nabla c(\mathbf{x}, t) \right] - \nabla \left(\mathbf{q}(\mathbf{x}) c(\mathbf{x}, t) \right) + \frac{q_s(\mathbf{x})}{\theta} c_s(\mathbf{x}, t) = \frac{\partial c(\mathbf{x}, t)}{\partial t} \tag{2}$$

Here, t denotes time; c is solute concentration; D is the (isotropic) dispersion coefficient; θ is effective porosity; c_s is solute concentration corresponding to q_s; and $\mathbf{q}(\mathbf{x}) = -\left(K(\mathbf{x}) / \theta \right) \nabla h(\mathbf{x})$ is an effective velocity associated with solute transport.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Numerical methods (e.g., finite differences or finite elements) are commonly employed to discretize Equations (1) and (2) that are then solved within a numerical MC context. The probability distribution of state variables of interest (e.g., heads or concentrations) is then evaluated at N nodes of an aptly designed numerical grid. Consistent with Section 1, we refer to the model corresponding to the numerical solution of the above equations as the Full System Model (FSM). When the domain is characterized by a large spatial extent and/or one is interested in exploring the system behavior across long temporal windows, performing numerical MC simulations relying on FSM is associated with a heavy computational burden. To circumvent this issue, we rely on the development and implementation of a Reduced-Order Model (ROM) strategy for solute transport. We note that in this study we employ ROM solely for solute transport because only limited computational costs are associated with the steady-state flow condition we consider, as opposed to simulating transport. Hereafter, we refer to numerical MC analyses grounded on ROM as ROMC.

2.2 Numerical Monte Carlo simulation framework for solute transport

2.2.1 Monte Carlo simulation setting for the Full System Model

We rely on a standard finite element method to solve the FSM described in Section 2.1. When considering a total simulation time T_s, we express the linear system associated with the numerical solution of solute transport through FSM within time interval $\left[t, t+\Delta t\right]$ as:

$$\mathbf{A}^i \mathbf{c}^i = \mathbf{F}^i \tag{3}$$

Here, superscript i refers the to the i^{th} MC realization (i = 1, …, N_{MC}, N_{MC} being the

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

total number of MC simulations) of FSM; $\mathbf{A}$ is the full-system stiffness matrix (of size N×N); $\mathbf{c}$ is the vector (of size N×1) of solute concentration values; and $\mathbf{F}$ is the stress vector (of size N×1) whose entries encompass source/sink terms and initial and boundary conditions.

2.2.2 Reduced-order Monte Carlo simulation framework

We construct a reduced-order model for solute transport by approximating the solution of solute concentration for the i^{th} MC realization of FSM. Consistent with the work of Xue and Xie (2007) and Pinnau (2008), one can approximate $\mathbf{c}^i$ as:

$$\mathbf{c}^i \approx \sum_{j=1}^{n} \alpha_j^i \mathbf{p}_j = \mathbf{P}\boldsymbol{\alpha}^i \tag{4}$$

Here, $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n]$ is a matrix (of size N×n, n being the dimension of the ROM) collecting the n nodal basis functions that are here obtained through a Proper Orthogonal Decomposition (POD) approach (see below); $\boldsymbol{\alpha}^i = \left[\alpha_1^i, \alpha_2^i, \cdots, \alpha_n^i\right]^T$ (T representing transpose) is a vector (of size n×1) of Fourier coefficients (Pinnau, 2008). Note that Equation (4) is different from a typical Karhunen-Loève expansion of $\mathbf{c}^i$ (i.e., $\mathbf{c}^i \approx \langle \mathbf{c} \rangle + \sum_{j=1}^{n} \alpha_j^i \mathbf{p}_j = \langle \mathbf{c} \rangle + \mathbf{P}\boldsymbol{\alpha}^i$, see Equation (11) in Li et al., 2013b). As we illustrate in Section 2.3, relying on Equation (4) enables straightforward (i) coding and (ii) compatibility with the iterative Ensemble Smoother (iES).

Substituting Equation (4) into Equation (3) and imposing the residual of the model equation associated with the approximated solution to be orthogonal to the projection space defined through $\mathbf{P}$ yields:

$$\mathbf{P}^T \mathbf{A}^i \mathbf{P}\boldsymbol{\alpha}^i \approx \mathbf{P}^T \mathbf{F}^i \tag{5}$$

Solving Equation (5) (which is a linear system of size n) yields $\boldsymbol{\alpha}^i$ for the i^{th}

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

MC realization of our ROMC strategy. Note that, when $n \ll N$, the computational effort required by our ROMC is much less than that of the standard MC.

The basis functions forming the entries of **P** are computed as from the leading eigenvectors (corresponding to the highest eigenvalues) of the covariance of totally N_{sn} numerical solutions (i.e., $\mathbf{c}^1$, $\mathbf{c}^2$, …, and $\mathbf{c}^{N_{sn}}$) through FSM. We point out that $N_{sn} = m \times N_t$, where m is the number of MC realizations which are arbitrarily chosen, each yielding $N_t = T_s / \Delta t$ numerical solutions of Equation (2), with uniform length of time step Δt. The leading eigenvectors are computed through the Singular Value Decomposition (SVD) approach, i.e.:

$$\mathbf{U\Lambda U}^T = svd\left(\mathbf{EE}^T\right) \tag{6}$$

where $\mathbf{E} = 1\big/\sqrt{N_{sn}}\left[\mathbf{c}^1, \mathbf{c}^2, \cdots, \mathbf{c}^{N_{sn}}\right]$; **U** (of size $N \times N$) is the left singular matrix whose j^{th} column is the j^{th} eigenvector of matrix $\mathbf{EE}^{\mathrm{T}}$ corresponding to the j^{th} singular value, λ_j; and $\mathbf{\Lambda} = diag\left(\left[\lambda_1, \lambda_2, \cdots, \lambda_N\right]\right)$ whose entries are ranked in descending order.

2.3 Iterative ensemble smoother

We denote by $\mathbf{m} = \left[Y_1, Y_2, \ldots, Y_N, \ln q_s, x_{1,q_s}, x_{2,q_s}\right]^T$ the vector (of size $P = N+3$) whose entries correspond to the uncertain model parameters (i.e., the log-conductivity, $Y = \ln K$, field) and flow rate and location of a pumping well. In case the pumping rate and location are known, then $\mathbf{m} = \left[Y_1, Y_2, \ldots, Y_N\right]^T$ and $P = N$. We further denote by $\mathbf{d} = \left[d_1, d_2, \ldots, d_O\right]^T$ the vector (of size O) of the observations (i.e., measured head and concentration values). To estimate **m**, we implement the iES (Luo and Bhakta, 2020; Xia et al., 2024):

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

$$\begin{cases} \mathbf{m}^{k+1} = \mathbf{m}^k + \underline{\underline{\mathbf{K}}}_{Gain}^k \Delta\mathbf{d}^k \\ \underline{\underline{\mathbf{K}}}_{Gain}^k = \underline{\underline{\mathbf{S}}}_m^k \left(\underline{\underline{\mathbf{S}}}_d^k\right)^T \left(\underline{\underline{\mathbf{S}}}_d^k \left(\underline{\underline{\mathbf{S}}}_d^k\right)^T + \gamma^k \underline{\underline{\mathbf{I}}}\right)^{-1} \quad \text{with} \quad \gamma^i = \xi^i \,\mathrm{trace}\left(\underline{\underline{\mathbf{S}}}_d^i \left(\underline{\underline{\mathbf{S}}}_d^i\right)^T\right) \Big/ O \,. \\ \Delta\mathbf{d}^k = g\left(\mathbf{m}^k\right) - \mathbf{d} \end{cases} \tag{7}$$

Here, superscript k is the index of the iteration step; matrices $\underline{\underline{\mathbf{S}}}_m^k = \left[\mathbf{m}_1^k - \overline{\mathbf{m}}^k, \cdots, \mathbf{m}_N^k - \overline{\mathbf{m}}^k\right] \Big/ \sqrt{N-1}$ (of size $P \times N$, where $\overline{\mathbf{m}}^k = \sum_{j=1}^{N} \mathbf{m}_j^k \Big/ N$) and $\underline{\underline{\mathbf{S}}}_d^k = \left[g\left(\mathbf{m}_1^k\right) - g\left(\overline{\mathbf{m}}^k\right), \cdots, g\left(\mathbf{m}_N^k\right) - g\left(\overline{\mathbf{m}}^k\right)\right] \Big/ \sqrt{N-1}$ (of size $O \times N$, where $g(\bullet)$ represents model operator being either FSM or ROM) collect the ensemble anomalies of parameters and simulated observations associated with the k^{th} iteration step; $\mathbf{I}$ is the identity matrix (of size $O \times O$); and ξ^k is an adaptive coefficient (Luo et al., 2015) associated with each iteration of the Levenberg-Marquardt (LM; Levenberg, 1944) algorithm. We set ξ^0 = 10 in our showcase application examples (see Section 3) and follow Luo and Bhakta (2020) to update its value for the remaining iteration steps.

In the case of $g(\bullet)$ representing the model operator of ROM, we note that the approximation of solute concentration relying on Equation (4) is compatible with the implementation of Equation (7). The degree of compatibility of ROM to iES is reduced when considering a typical Karhunen-Loève expansion of $\mathbf{c}^i$ (i.e., $\mathbf{c}^i \approx \langle\mathbf{c}\rangle + \sum_{j=1}^{n} \alpha_j^i \mathbf{p}_j = \langle\mathbf{c}\rangle + \mathbf{P}\boldsymbol{\alpha}^i$). This is related to the observation that $\langle\mathbf{c}\rangle$ evolves with time and needs to be evaluated at each time step. This, in turn, implies that m numerical solutions of solute concentration through FSM need to be obtained to evaluate $\langle\mathbf{c}\rangle$ at every outer iteration of iES. Hence, computational advantages of employing ROM are reduced while coding complexity increases. When approximating solute concentration via Equation (4), we only obtain m numerical

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

solutions of solute concentration through FSM at the first outer iteration of iES. Leading eigenvectors are computed upon relying on these solutions and are then stored. The Fourier coefficients $\boldsymbol{\alpha}^i$ associated with time interval $\left[t, t+\Delta t\right]$ for each MC realization starting from the second outer iteration of iES are obtained solely through solving Equation (5).

When implementing the LM algorithm during optimization, we set both the inner and outer iteration numbers equal to 10 (see also Luo and Bhakta, 2020). Additionally, a stopping criterion $\left(\delta_{k-1}-\delta_k\right)/\delta_{k-1}\times 100\% \leq 10^{-6}$ (where $\delta_k=\frac{1}{N}\sum_{j=1}^{N}\left\{\left(\mathbf{d}_j^k-g\left(\mathbf{m}_j^k\right)\right)^T \underline{\underline{\mathbf{C}}}_d^{-1}\left(\mathbf{d}_j^k-g\left(\mathbf{m}_j^k\right)\right)\right\}$), is set.

2.4 Computational cost

We denote by iES_FSM and iES_ROM the approaches associated with coupling the iES with FSM and ROM, respectively. The total number of MC realizations is denoted as N_{MC}. Neglecting the computational cost of the inner iterations and assuming iES comprises N_{out} outer iterations, the main computational costs of either method can be divided into two components, corresponding to forecast and analysis step (Table 1), respectively. In the forecast step, a number of ($N_{out}+1$) MC simulations for groundwater flow and solute transport are required. Otherwise, Equation (7) is evaluated N_{out} times in the analysis step. The steady-state groundwater flow is solved through the FSM in both iES_FSM and iES_ROM, with a computational cost of order $O\left(N^3 N_{MC}\right)$. The main computational cost for the N_{MC} FSM-based MC realizations of solute transport at a single time step in iES_FSM is

© Author(s) 2025. CC BY 4.0 License.

$O\left(N^3 N_{MC}\right)$, while being $O\left(\left(sN + N^2\right) N_{MC}\right)$ (where $s \approx 7$ or ≈ 15 in two and three dimensions, respectively) for iES_ROM. These computational efforts correspond to the projection of the full-system stiffness matrix onto the reduced-order space of the system state (i.e., solute concentration). Computational costs associated with solving Equation (7) coincide for both approaches and are here denoted as C_8. We further note that, with reference to iES_ROM, the N_{sn} solutions of solute concentration obtained through FSM (associated with a computational cost of order $O\left(N^3 N_{sn}\right)$) and the basis functions obtained through SVD (with a computational cost of order $O\left(nN_{sn}^2\right)$) are calculated only once and stored. When the grid mesh employed is large or the simulation time is long, computational savings through iES_ROM compared with iES_FSM become significant.

3. Exemplary scenarios

We consider a two-dimensional computational domain of size 4 × 2 to simulate a synthetic sandbox-scale experiment where (non-reactive) solute transport under steady-state flow is considered (see Fig. 1). Here and hereafter, all quantities are given in consistent (length/mass/time) units. Concerning groundwater flow, the left and right sides of the domain are associated with constant head boundary conditions with H = 3 and 2, respectively. The top and bottom sides correspond to boundary conditions of no flow. A pumping well with an unknown pumping rate and location is considered in the setting. A fixed concentration boundary is set at point (0, 1) (see red triangle in Fig. 1) with a constant concentration of 100, while the initial concentration across the domain is set to zero. We use the standard finite element method to obtain the numerical

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

solutions of head and concentration. The numerical mesh adopted comprises 41×21 = 861 nodes and 1,600 triangle elements. A uniform time step of 1 day is considered for a total simulation time of 10 days.

The logarithm of conductivity (Y = ln$K$) is considered as a spatially heterogeneous (correlated) Gaussian random field with an exponential covariance function (C_Y) given by:

$$C_Y = \sigma_Y^2 \exp\left(-\left(\frac{d_{x_1}}{\lambda_{x_1}} + \frac{d_{x_2}}{\lambda_{x_2}}\right)\right) \tag{8}$$

where σ_Y^2 is the variance of Y; d_{x_i} (i = x, y) is separation (lag) distance between two given points in the i-direction; λ_{x_i} (with i = x, y) is the correlation length of Y in the i-direction. The corresponding mean of Y is denoted as μ. The initial ensemble of Y fields is synthetically generated through the well-known and widely tested GSLIB suite (Deutsch and Journel, 1998) upon setting λ_{x_1} and λ_{x_2} equal to 1.0 and 0.5, respectively. The reference Y field (Fig. 1a) is generated upon setting μ = 0.8, σ_Y^2 = 1.0, λ_{x_1} = 1.0, and λ_{x_2} = 0.5.

The pumping rate (i.e., q_s), x_1 and x_2 -coordinates (denoted as x_{1,q_s} and x_{2,q_s}, respectively) of the pumping location are considered to be random variables, each associated by a Gaussian distribution. The gray zone in Fig. 1b encompasses the possible locations where a pumping well is operating. The initial collection (ensemble) of values of q_s, x_{1,q_s}, and x_{2,q_s} and their reference counterparts are sampled from Gaussian distributions characterized by mean (standard deviation) equal to 0.50 (0.25), 1.00 (0.25), and 1.00 (0.25), respectively. These settings ensure that the randomly

© Author(s) 2025. CC BY 4.0 License.

generated samples of x_{1,q_s} and x_{2,q_s} are mostly within the coordinate ranges indicated by the gray zone in Fig. 2b. Reference values are q_s = 1.03, x_{1,q_s} = 1.38, and x_{2,q_s} = 1.40 (see Fig. 1b, red cross symbol). Figure 1c depicts the simulated head field associated with the reference conductivity field, pumping rate, and location. Figure 1d depicts simulated concentrations at the final simulation time. Observations, including (steady-state) head and solute concentration at each time step, are collected at a number (denoted as N_m) of monitoring wells distributed across the aquifer according to some pre-defined patterns (Fig. 1b-d). Each measurement is taken as the sum of the simulated head (or concentration) and a white noise with zero mean and standard deviation equal to σ_{obs}.

To explore the potential of iES_ROM, several showcases are designed to highlight key features of interest. Four groups of test cases (TCs) are designed and organized as detailed in the following (see also Table 1).

- **Group A**. It includes twelve TCs (i.e., TC1-TC12), enabling us to compare performances of iES_FSM and iES_ROM associated with diverse values of *n* when the pumping rate and locations are either known (TC1-TC6) or unknown (TC7-TC12). The dimension of the ROM is considered equal to {5, 10, 15, 20, 25, 30}, these values being consistent with those most commonly analyzed in previous studies (Pasetto et al., 2014; Xia et al., 2020, 2025).
- **Group B**. It includes four TCs (i.e., TC6 and TC13-TC15), enabling us to compare the performances of iES_FSM and iES_ROM with the largest value of *n* analyzed (i.e, *n* = 30) and considering diverse values of N_{MC}

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

corresponding to {30, 100, 500, 10,000}. The latter are values of N_{MC} commonly tested in previous studies (Chen and Zhang, 2006; Xia et al., 2021, 2024).

- **Group C**. It includes five TCs (i.e., TC6 and TC16-TC19), designed to analyze the ability of iES_ROM to cope with diverse quality and quantity of available measurements. Performances of iES_FSM and iES_ROM are also compared when σ_{obs} = {0.001, 0.01, 0.1} and the number of observation locations corresponds to a value selected from {9 (Fig. 1b), 18 (Fig. 1c), 55 (Fig. 1d)}.
- **Group D**. It includes five TCs (i.e., TC6 and TC_20-TC23), enabling us to study the effect of μ and σ_Y^2 of the initial ensemble of Y on the accuracies of estimates of conductivity and pumping rate and well location through iES_FSM and iES_ROM. Values of μ and σ_Y^2 of the initial ensemble of Y fields are selected from {-0.5, 1.2, 2.0} and {0.01, 1.0, 2.0}, respectively.
- **Group E**. It includes six TCs (i.e., TC6 and TC24-TC28), with the aim of investigating the effect of N_{sn} on the accuracies of the estimation of conductivity and well pumping rate and location through iES_ROM and on computation time requirements. Values of N_{sn} in TC24-TC28 and TC6 are equal to 30, 100, 300, 500, 1,000, and 10,000, respectively.

Note that, without specified otherwise, default settings for the above mentioned TCs correspond to TC6 which is designed with n = 30, N_{MC} = 10,000, N_{sn} = 10,000,

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

N_m = 55, σ_{obs} = 0.01, and values of μ and σ_Y^2 of the initial ensemble of Y equal to 1.2 and 1.0, respectively. Except for TC8-TC12, the source/sink term is associated with uncertainty.

To quantify the accuracy of conductivity estimates through iES_ROM and iES_FSM, we consider absolute error between estimated and reference values of Y (denoted as E_Y) and estimate of the standard deviation (denoted as S_Y) which are defined as:

$$E_Y = \frac{1}{N}\sum_{i=1}^{N}\left|\left\langle Y_i \right\rangle^{est} - Y_i^{ref}\right|; \qquad S_Y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\sigma_{Y,i}^2\right)^{est}} \tag{9}$$

where $\left\langle Y_i \right\rangle^{est}$, $\left(\sigma_{Y,i}^2\right)^{est}$, and Y_i^{ref} denote estimated (ensemble) mean and variance, and reference value of Y at the i^{th} cell of the numerical grid, respectively.

Absolute errors and estimates of the standard deviations of $\ln q_s$, x_{1,q_s}, and x_{2,q_s} are employed to quantify the accuracy of the estimate of the pumping rate and well location:

$$E_{q_s} = \left|\left\langle \ln q_s \right\rangle^{est} - \ln q_s^{ref}\right|; \qquad E_{x_1} = \left|\left\langle x_{1,q_s} \right\rangle^{est} - x_{1,q_s}^{ref}\right|; \qquad E_{x_2} = \left|\left\langle x_{2,q_s} \right\rangle^{est} - x_{2,q_s}^{ref}\right| \tag{10}$$

where $\left\langle \ln q_s \right\rangle^{est}$, $\left\langle x_{1,q_s} \right\rangle^{est}$, and $\left\langle x_{2,q_s} \right\rangle^{est}$ indicate estimated (ensemble) mean values of $\ln q_s$, x_{1,q_s}, and x_{2,q_s} respectively; and q_s^{ref}, x_{1,q_s}^{ref}, and x_{2,q_s}^{ref} are the reference values of q_s, x_{1,q_s}, and x_{2,q_s}, respectively. Estimates of the standard deviations of $\ln q_s$, x_{1,q_s}, and x_{2,q_s} are:

$$S_{q_s} = \sqrt{\left(\sigma_{\ln q_s}^2\right)^{est}}; \qquad S_{x_1} = \sqrt{\left(\sigma_{x_{1,q_s}}^2\right)^{est}}; \qquad S_{x_2} = \sqrt{\left(\sigma_{x_{2,q_s}}^2\right)^{est}} \tag{11}$$

where $\left(\sigma_{\ln q_s}^2\right)^{est}$, $\left(\sigma_{x_{1,q_s}}^2\right)^{est}$, and $\left(\sigma_{x_{2,q_s}}^2\right)^{est}$ denote estimated (ensemble) variances of $\ln q_s$, x_{1,q_s}, and x_{2,q_s}, respectively.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

As an additional metric, we then rely on the average absolute difference between available data and model results:

$$E_{obs} = \frac{1}{O}\sum_{i=1}^{O}\left|\left\langle d_i \right\rangle^{up} - d_i^{ref}\right| \tag{12}$$

where $\left\langle d_i \right\rangle^{up}$ and d_i^{ref} correspond to the (updated) result of the simulation process and its reference observed counterpart at the i^{th} sampled location, respectively.

4. Results and discussion

4.1 Impact of the dimension of the reduced-order model (Group A)

Figure 2 depicts E_Y (Fig. 2a), S_Y (Fig. 2b), and E_{obs} (Fig. 2c) versus the number of outer iterations for test cases (TCs) 1-6 obtained through iES_ROM and iES_FSM, when well pumping rate and location are uncertain. Note that results obtained through iES_FSM are independent of *n* (and are identical among TCs 1-6) and are taken as references. Percentage differences (denoted as ΔE_Y) between the values of E_Y obtained through iES_ROM and iES_FSM are depicted in Fig. 2d. Corresponding results associated with percentage differences between values of S_Y (ΔS_Y) and of E_{obs} (ΔE_{obs}) are depicted in Fig. 2e and 2f, respectively.

Values of E_Y, S_Y, and E_{obs} obtained at the end of the iteration procedure through iES_ROM generally decrease with *n*. When *n* = 25 or 30, the values of E_Y and S_Y based on iES_ROM tend to approach their counterparts obtained through iES_FSM. The latter generally correspond to the lowest values across TCs 1-6. These findings are consistent with the observation (Xia et al., 2020; 2025) that accuracy of ROM for *n* = 30 and FSM are very similar for solute transport. They are also in line with the results of Li et al. (2013b), who documented a high degree of correlation

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

between simulated concentrations provided by their ROM and FSM for non-reactive transport. As expected, the solution accuracy of ROM increases with n. When n approaches the dimension of the FSM (i.e., the representativeness of the basis functions is essentially guaranteed), the ROM-based solution approaches its FSM-based counterpart.

Figure 3 depicts E_Y (Fig. 3a), S_Y (Fig. 3b), and E_{obs} (Fig. 3c) for TCs 7-12 obtained through iES_ROM and iES_FSM when the well characteristics are deterministically known. Similar to above, results obtained through iES_FSM are identical among TCs 7-12 and are taken as reference. Values of ΔE_Y, ΔS_Y, and ΔE_{obs} are depicted in Fig. 3d, 3e, and 3f, respectively.

Consistent with what one can observe in Fig. 2, values of E_Y, S_Y, and E_{obs} obtained at the end of the iteration procedure for TCs 7-12 through iES_ROM generally decrease with n. Except for the cases where n = 5 or 10 (corresponding to low solution accuracy of ROM), values of E_Y, S_Y, and E_{obs} for TCs 9-12 based on either iES_ROM or iES_FSM are lower than their counterparts related to TCs 3-6. These results suggest that the accuracy of conductivity estimates is lower when q_s is uncertain compared to the case where q_s is deterministic.

Figure 4 depicts the values of E_{x_1} (Fig. 4a), E_{x_2} (Fig. 4b), E_{q_s} (Fig. 4c), S_{x_1} (Fig. 4d), S_{x_2} (Fig. 4e), and S_{q_s} (Fig. 4f) versus the number of outer iterations for TCs 1-6 obtained through iES_ROM and iES_FSM. Values of E_{x_1}, E_{x_2}, and E_{q_s} obtained through iES_ROM approach their iES_FSM-based counterparts as n increases. This is consistent with the observation that increasing n improves the

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

accuracy of the ROM-based solution (see also Li et al., 2013b), therefore enhancing the accuracy of the identification of the well attributes.

Figure 5 depicts the estimated (ensemble) Y fields for TCs 1-6 obtained through iES_ROM and iES_FSM, together with their reference Y field. The white circle and cross symbols in Fig. 5 denote the estimated and reference locations of the pumping well, respectively. As n increases, the estimated Y field obtained through iES_ROM (Fig. 5a-5f) approaches its iES_FSM-based counterpart and the reference Y field (Fig. 5h). The accuracy of the iES_ROM-based estimate of the location of the pumping well generally increases with n, consistent with the nature of the findings illustrated in Fig. 4. Figure 6 depicts the estimated (ensemble) Y variance fields for TCs 1-6 based on iES_ROM and iES_FSM. The white circle and cross symbols therein denote the identified and reference locations of the pumping well, respectively. These results show that the variance of Y is overestimated when n is small. This is related to the observation that small values of n correspond to large modeling errors (i.e., low solution accuracy) of ROM (as also seen in Li et al. (2013b) and Pasetto et al. (2017)). The latter, in turn, imprint the low accuracy of conductivity estimates (see Fig. 5a in the case of n = 5) and yield overestimated values for the variance of Y (see Fig. 6a).

Figure 7 depicts the empirical probability density function (PDF) of x_{1,q_s}, x_{2,q_s}, and $\ln q_s$ at the end of the iteration procedure for TCs 1, 2, 4, and 6 as obtained through iES_ROM and iES_FSM, together with their counterparts associated with initial guess (black solid) and reference values (black dashed). One can observe that large values of n yield high accuracy for x_{1,q_s} and x_{2,q_s} estimates, as visually

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

indicated by the compact supports associated with the empirical PDFs of x_{1,q_s} (Fig. 7a) and x_{2,q_s} (Fig. 7b). The accuracy of the estimate of q_s is already acceptable when $n = 5$.

As an additional element, we explore the way the choice of the value of n impacts the local PDFs of hydraulic head and solute concentration. We do so upon considering the results associated with three reference points (i.e., I, II, and III in Fig. 1d) that are aligned in the direction of the mean groundwater flow. Figure 8 depicts the (sample) PDFs of (hydraulic) head at these three selected locations (Figs. 8a-8c) obtained through iES_ROM and iES_FSM at the end of the iteration procedure for TCs 1, 2, 4, and 6. Black solid lines included therein indicate reference head values. Note that the PDFs stemming from iES_FSM peak at values very close to their reference counterparts. Hence, the corresponding empirical PDFs are considered as reference. The logarithm absolute difference (ΔPDF, evaluated as the pointwise log-ratio of the densities and corresponding to a local measure of relative likelihood between two empirical PDFs) between the PDFs of the head at points I-III obtained through iES_ROM based on diverse values of n and their counterpart based on iES_FSM are also shown in Figs. 8d-8f, respectively. One can see that a large value of n (e.g., n = 30 for TC6) corresponds to high accuracy of the PDF of head, as quantified through a low value of ΔPDF. Although the head solution is obtained by solving FSM, the accuracy of the conductivity estimate is impacted by n. The latter, therefore, impacts the accuracy of heads. Fig. 9 depicts results related to solute concentration. As expected, the PDFs stemming from iES_FSM peak at values very close to their reference counterparts also in this case. Consistent with Fig. 8, a large value of n (e.g., 30 for TC6) corresponds to high accuracy in the delineation of the PDF of solute concentration.

As a complement to these results, values of the Kullback-Leibler Divergence (KLD) between the (sample) PDFs of head at the three reference points at the last outer iteration obtained through iES_FSM (h_{FSM}) and iES_ROM (h_{ROM}) with n = 5 (TC1), 10 (TC2), 20 (TC4), and 30 (TC6) are listed in Table S1 (see supplementary information). We recall that values of KLD($h_{\mathrm{ROM}}||h_{\mathrm{FSM}}$) (or (KLD($h_{\mathrm{FSM}}||h_{\mathrm{ROM}}$)) quantify (in a global sense) information loss when using h_{FSM} (h_{ROM}) to approximate h_{ROM} (h_{FSM}). Values of KLD($h_{\mathrm{ROM}}||h_{\mathrm{FSM}}$) generally increase with n. This indicates that the difference between PDFs of h_{ROM} and h_{FSM} decrease as n increases. While the highest values of KLD($h_{\mathrm{FSM}}||h_{\mathrm{ROM}}$) correspond to n = 5, no clear decreasing trends with increasing n are observed. Furthermore, the difference between KLD($h_{\mathrm{ROM}}||h_{\mathrm{FSM}}$) and KLD($h_{\mathrm{FSM}}||h_{\mathrm{ROM}}$) generally decreases as n increases. This is related to the observation that the accuracy of ROM tends to increase as the dimension of the reduced-order model increase. Values of KLD between the empirical PDFs of solute concentrations at the three selected reference points at the last outer iteration obtained through iES_FSM (c_{FSM}) and iES_ROM (c_{ROM}) with n = 5 (TC1), 10 (TC2), 20 (TC4), and 30 (TC6) are listed in Table S2 (see supplementary information).

4.2 Effect of the ensemble size (Group B)

Figure 10 depicts iES_ROM- and iES_FSM-based values of E_Y (Fig. 10a), S_Y (Fig. 10b), and E_{obs} (Fig. 10c) versus the number of outer iterations for TCs 6 and

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

13-15. Values of E_Y and E_{obs} decrease as the ensemble size N_{MC} increases (while the value of S_Y increases) regardless of the approach employed. With reference to TC13, we note that when N_{MC} = 30 the values of E_Y decrease during the course of the first outer iterations to then increase during the last outer iterations, values of S_Y dropping rapidly during the iteration procedure, regardless of the approach employed. This phenomenon is typically linked to the occurrence of filter inbreeding caused by a limited ensemble size (Chen and Zhang, 2006; Xia et al., 2018; 2024). Values of E_Y and S_Y for TCs 6 and 13-15 obtained through iES_ROM are overall similar to those associated with iES_FSM. The iES_ROM-based value of E_{obs} obtained at the end of the iteration procedure for a given TC is typically larger than its iES_FSM-based counterpart. This is linked to the observation that the limited system dimension of ROM induces low accuracy of concentrations and (possibly) heads due to low accuracy of conductivity estimates, pumping rate, and well locations.

Figure 11 depicts the values of E_{x_1} (Fig. 11a), E_{x_2} (Fig. 11b), E_{q_s} (Fig. 11c), S_{x_1} (Fig. 11d), S_{x_2} (Fig. 11e), and S_{q_s} (Fig. 11f) versus the number of outer iterations for TCs 6 and 13-15 obtained through iES_ROM and iES_FSM. When increasing N_{MC}, values of E_{x_1}, E_{x_2}, and E_{q_s} obtained through either iES_ROM or iES_FSM do not show a clear trend. Values of S_{x_1}, S_{x_2}, and S_{q_s} generally increase with N_{MC}, a result that is consistent with the findings encapsulated in Fig. 10b. Similar findings are also documented by Xu and Gómez-Hernández (2018, their Fig. 17), who show that, when considering joint identification of contaminant sources and hydraulic conductivities, the accuracy of estimates of key attributes characterizing contaminant

© Author(s) 2025. CC BY 4.0 License.

sources does not necessarily improve after some time and as data assimilation progresses. We further note that jointly estimating conductivity and identifying source/sink term attributes (in terms of flow rate and location) is associated with a highly nonlinear optimization process. Hence, the accuracies of location and pumping rate estimation through iES_FSM are not always higher than those stemming from iES_ROM in terms of the values of the metrics employed (i.e., E_{x_1}, E_{x_2}, and E_{q_s}).

Figures 12 depicts the estimated (ensemble mean) Y fields for TCs 6 and 13-15 obtained through iES_ROM and iES_FSM. Figure 13 depicts the associated Y variance fields for TCs 6 and 13-15 obtained through iES_ROM and iES_FSM. The white (black) circle and cross symbols in Fig. 12 (or Fig. 13) represent the identified and the reference locations of the pumping well, respectively. Visual comparison of Fig. 12 and Fig. 5h suggests that the estimated Y fields rendered through an ensemble size N_{MC} = 100 (i.e., TC14) obtained through iES_ROM and iES_FSM are the closest ones to the reference Y field. Nevertheless, jointly analyzing Figs. 10a, 12, and 13 reveal that the estimated Y field corresponding to N_{MC} = 10,000 (TC6) obtained through iES_ROM is the one most closest to the reference Y field in terms of E_Y (= 0.41). Additionally, the identified and reference locations of the pumping well obtained through either iES_ROM or iES_FSM are close to each other, thus supporting the capability of both approaches to identify the well location.

4.3 Effect of quality and available number of observations (Group C)

Table 2 lists values of E_Y, S_Y, E_{obs}, E_{x_1}, E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, and S_{q_s} at the end of iteration procedure for TC16 (characterized by σ_{obs} = 0.001), TC6 (σ_{obs} =

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

0.01), and TC17 (σ_{obs} = 0.1) obtained through iES_ROM and iES_FSM. Values of E_Y, S_Y, E_{obs}, E_{x_1}, and E_{q_s} generally increase as the quality of observations deteriorates, i.e., σ_{obs} increasing from 0.001 to 0.1. These results are also consistent with prior findings by Xia et al. (2018) according to which accuracy of conductivity estimates increases as the quality of observations improves. Values of E_{x_2} obtained through iES_ROM and iES_FSM do not monotonically decrease as σ_{obs} decreases. This is typically related to the strong nonlinear nature associated with the optimization process (see also Xu and Gómez-Hernández, 2018).

Figure 14 depicts iES_ROM- and iES_FSM-based values of E_Y (Fig. 14a), S_Y (Fig. 14b), and E_{obs} (Fig. 14c) versus the number of outer iterations for TCs 6 and 18-19. Values of E_Y (or S_Y) for TCs 18 (where the number of monitoring wells is N_m = 9), 19 (N_m = 18), and 6 (N_m = 55) obtained through iES_ROM are similar to their iES_FSM-based counterparts and decrease as N_m increases. Values of E_{obs} obtained through iES_FSM decrease as N_m increases, while iES_ROM-based results do not display a clear trend with N_m. This result may be attributed to the fact that, while increasing the number of monitoring wells enhances the amount of information available for estimating hydraulic conductivity, errors introduced through model reduction influence the evolution of the solute concentration mismatch between observations and simulations during the iterative calibration process.

Figure 15 depicts the values of E_{x_1} (Fig. 15a), E_{x_2} (Fig. 15b), E_{q_s} (Fig. 15c), S_{x_1} (Fig. 15d), S_{x_2} (Fig. 15e), and S_{q_s} (Fig. 15f) versus the number of outer iterations for TCs 6 and 18-19 obtained through iES_ROM and iES_FSM. Values of

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

E_{x_1} (E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, S_{q_s}, or S_Y) for TCs 18 (for a number N_m = 9 of monitoring wells), 19 (N_m = 18), and 6 (N_m = 55) obtained through either iES_ROM or iES_FSM decrease as N_m increases. Values of the same metric (i.e., E_{x_1}, E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, or S_{q_s}) obtained through iES_ROM and iES_FSM are overall close to each other.

Figure 16 depicts the empirical PDF of x_{1,q_s}, x_{2,q_s}, and $\ln q_s$ at the end of the iteration procedure for TCs 18, 19, and 6 obtained through iES_ROM and iES_FSM, together with their reference counterparts (black dashed lines). One can observe that increasing N_m leads to improved accuracy of the identification of pumping well attributes, as suggested by the reduced support and location of the peaks of the PDFs of x_{1,q_s} (Fig. 16a), x_{2,q_s} (Fig. 16b), and $\ln q_s$ (Fig. 16c) obtained through either iES_ROM or iES_FSM and observed as N_m varies from 9 to 55. On the basis of these results, it is hard to tell which approach provides higher accuracy of pumping well identification, solely in terms of Fig. 16. To complement these findings, Table S3 (see supplementary information) lists the values of KLD between the empirical PDFs of x_{1,q_s} (x_{2,q_s}, or $\ln q_s$) obtained through iES_FSM (denoted as p_{FSM}) and iES_ROM (denoted as p_{ROM}) with n = 30, considering N_m = 9 (TC18), 18 (TC19), and 55 (TC6), respectively. Values of KLD($p_{j\mathrm{ROM}}||p_{j\mathrm{FSM}}$) (with j = x_{1,q_s}, x_{2,q_s}, and $\ln q_s$ show an overall decreasing trend as N_m increase, while KLD($p_{j\mathrm{FSM}}||p_{j\mathrm{ROM}}$) consistently decreases with N_m. These results are consistent with the observation that increasing the number of monitoring wells improves the accuracy of conductivity estimates (as also seen by Tong et al. (2010) and Xia et al. (2018)) as well as pumping

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

rate and well location through both approaches, thus, in turn, reducing discrepancies between the corresponding PDFs.

4.4 Effect of the mean and variance of the initial ensemble of *Y* (Group D)

Table 3 lists the values of E_Y, S_Y, E_{obs}, E_{x_1}, E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, and S_{q_s} at the end of the iteration procedure for TCs 20 (characterized by a mean μ = -0.5 of the initial ensemble of *Y*), 6 (μ = 1.2), and 21 (μ = 2.0) obtained through iES_ROM and iES_FSM. We recall that the mean value employed to generate the reference *Y* field is equal to 0.8. When the discrepancy between μ and the mean value of the reference *Y* field increases, the error metrics employed display an overall increase, E_{x_1} and E_{q_s} constituting notable exceptions. This finding is consistent with the behavior documented by Xia et al. (2024) who considered two correlation-based localization approaches to assess conductivity estimation accuracy with respect to the mean of the initial ensemble of *Y*. Overall, this result is related to the highly nonlinear nature of the optimization process (see also Sections 4.2 and 4.3) stemming from the joint estimation of conductivities and pumping well attributes.

Table 4 lists the values of E_Y, S_Y, E_{obs}, E_{x_1}, E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, and S_{q_s} at the end of the iteration procedure for TCs 22 (characterized by a variance σ_Y^2 = 0.01 of the initial ensemble of *Y*), 6 (σ_Y^2 = 1.0), and 23 (σ_Y^2 = 2.0) obtained through iES_ROM and iES_FSM. We recall that the reference *Y* field is characterized by a unit variance. The values of E_Y and E_{x_2} obtained through both approaches increase as the discrepancy between σ_Y^2 and the variance of the reference *Y* field increases. The values of E_{obs}, E_{x_1}, and E_{q_s} obtained through both approaches generally increase with σ_Y^2. Similarly, values of metrics employed to quantify variability of the final ensemble of realizations (i.e., S_Y, S_{x_1}, S_{x_2}, and S_{q_s}) consistently increase with σ_Y^2.

A joint analysis of the results illustrated in Sections 4.1, 4.2, and 4.3 suggests that E_Y and S_Y provided by both approaches show a consistent behavior as a function of the key feature of interest. Otherwise, the response of the metrics associated with the pumping well attributes provided by both approaches reflects the enhanced nonlinearity of the associated optimization process. Additionally, the accuracy of the conductivity estimate possibly contributes more to the minimization of the objective function than that of pumping well identification. Additionally, the values of the metrics in Sections 4.1, 4.2, and 4.3 provided by the two approaches are generally consistent with each other, thus supporting the representativeness of the iES_ROM-based results.

4.5 Effect of the snapshot size (Group E)

Table 5 lists percentage differences of the values of the performance metrics considered (i.e., E_Y, S_Y, E_{obs}, E_{x_1}, E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, and S_{q_s}) at the end of the iteration procedure for TCs 24-28 obtained through iES_ROM, considering their counterparts through TC6 as references. These results show that the values of E_Y and S_Y systematically decrease as N_{sn} increases from 30 to 1,000, while the other metrics display an overall decreasing pattern. This is related to the observation that a larger snapshot size corresponds to a higher accuracy of basis functions (Pasetto et al., 2014). Otherwise, it is worth noting that snapshots are evaluated only once throughout

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

the entire data assimilation processes, thus resulting in a limited computational cost.

The CPU required time for running TC28 is about 13 minutes (using a processor 13th Gen Intel(R) Core(TM) i7-13700K 3.40 GHz with 32 GB RAM). CPU times for running TC6 through iES_ROM and iES_FSM are about 28 and 122 minutes, respectively. The CPU time to complete TC6 upon relying on iES_FSM is about 9 times the CPU time for TC28 through iES_ROM, while percentage differences associated with E_Y and S_Y are 0.50% and 0.21%, respectively. CPU time savings can become more pronounced during data assimilation for a groundwater system of large size, due to the higher memory requirements of iES_FSM for storing and computing large-dimensional vectors and matrices as compared to iES_ROM. These results support the ability of iES_ROM to estimate conductivity and identify the uncertain features of a pumping well under the conditions analyzed.

5. Conclusions

This study addressed the joint estimation of (uncertain, spatially heterogeneous) hydraulic conductivities and attributes (location and flow rate) of a pumping well in a two-dimensional confined aquifer in the presence of (non-reactive) solute transport taking place across a steady-state flow field. Our analyses rest on an iterative Ensemble Smoother (iES) coupled with a Reduced-Order Model (ROM) for solute transport (the overall strategy being denoted as iES_ROM). The ROM is constructed via Proper Orthogonal Decomposition (POD), using basis functions derived from the numerical solutions of the Full System Model (FSM) over the entire simulation period. The pumping well is characterized by its spatial coordinates (x_{1,q_s}, x_{2,q_s}) and a

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

constant pumping rate q_s. The ROM can achieve a solution accuracy similar to that of the FSM, while significantly reducing computational demands. Notably, as stated above, the basis functions are computed only once throughout the iES_ROM iteration process, thus further enhancing efficiency. As a benchmark, the traditional iES approach relying on the FSM (termed iES_FSM) is also implemented to estimate conductivity and identify well attributes.

To assess the performance and robustness of the proposed iES_ROM approach, twenty-eight test cases (TCs 1-28) are designed and structured according to five categories (Groups A-E; Section 3), each targeting different influencing factors. These include the dimension of the reduced-order model (n), ensemble size (N_{mc}), standard deviation of the white noise representing measurement error (σ_{obs}), number of monitoring wells (N_m), mean (μ) and variance (σ_Y^2)of the initial log-conductivity field, and snapshot size (N_{sn}). The performance of iES_ROM is systematically compared with that of iES_FSM using nine evaluation metrics, encompassing the absolute error (E_Y; Equation (9)) and estimated standard deviation (S_Y; Equation (9)) between estimated and reference values of Y; the absolute errors and estimated standard deviations of the pumping well coordinates and rate (Equation (10)); and the average absolute difference between simulated and reference observations (E_{obs}; Equation (12)).

Our work leads to the following major conclusions.

1. Both iES_ROM and iES_FSM yield accurate estimates of hydraulic conductivity distributions and identify the pumping well attributes across a

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

wide range of tested conditions, including variations in model dimension, ensemble size, measurement noise, number of monitoring wells, and statistical properties of the initial ensemble.

2. The iES_ROM approach achieves estimation accuracy similar to that of iES_FSM when using a moderate reduced-order dimension (n = 25 or 30). Otherwise, relying on a small dimension (e.g., n = 5) yields filter divergence due to unaccounted model errors. Increasing n effectively mitigates this issue and enhances the stability of the iES_ROM performance.
3. When hydraulic conductivity and pumping well attributes are jointly estimated, both iES_ROM and iES_FSM exhibit a slight reduction in the accuracy of conductivity estimates compared to scenarios where only conductivity is estimated. This trend is reflected in the values of E_Y, S_Y, and E_{obs} across TCs 1-12. Under such joint estimation, results in terms of E_Y, S_Y, and E_{obs} with respect to different influencing factors remain of acceptable quality for both iES_ROM and iES_FSM, consistent with the patterns observed in conductivity-only estimation. The behaviors of the remaining performance metrics are mutually consistent and within acceptable ranges, although somewhat less orderly.
4. Relying on the iES_ROM approach yields an accuracy similar to that of iES_FSM in estimating hydraulic conductivity and identifying pumping well attributes for both moderate (N_{sn} = 500 or 1000) and large (N_{sn} = 10000) ensemble sizes. This result supports its robustness with respect to ensemble

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

size selection.

5. In terms of computational efficiency, iES_ROM yields substantial time savings compared to iES_FSM. For instance, with N_{sn} = 500 and n = 30, the CPU times for iES_ROM and iES_FSM are approximately 28 and 122 minutes, respectively (i.e., iES_FSM requires a computation time that is about nine times longer while yielding similar estimation accuracy).

Author contributions. All authors contributed to the preparation of the manuscript.

Acknowledgments. This work was supported by the National Nature Science Foundation of China (Grant No. 42002247), Nature Science Foundation of Fujian Province, China (Grant No. 2025J01529; 2025J08248), and Opening Fund of Key Laboratory of Geohazard Prevention of Hilly Mountains, Ministry of Natural Resources (FJKLGH2024K008). M.R. acknowledges funding from the National Recovery and Resilience Plan (NRRP), mission 4 component 2 investment 1.4 - call for tender no. 3138 of 16 December 2021, rectified by decree no. 3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union - NextGenerationEU, project code CN_00000033, concession decree no. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP D43C22001250001, project title "National Biodiversity Future Center - NBFC".

References

Asher, M.J., Croke, B.F.W., Jakeman, A.J., Peeters, L.J.M., 2015. A review of surrogate models and their application to groundwater modeling. Water Resour. Res. 51, 5957–5973.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Ballio, F., Guadagnini, A., 2004. Convergence assessment of numerical Monte Carlo simulations in groundwater hydrology. Water Resour. Res. 40, W04603.

Boyce, S.E., Nishikawa, T., Yeh, W.W.G., 2015. Reduced order modeling of the newton formulation of modflow to solve unconfined groundwater flow. Adv. Water Resour. 83, 250–262.

Chen, Y., Zhang, D., 2006. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Adv Water Resour., 29(8): 1107-22.

Chen, Y., Oliver, D. S., 2013. Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. Computational Geosciences, 17(4): 689-703.

Chen, Z., Jaime Gomez-Hernandez, J., Xu, T., Zanini, A., 2018. Joint identification of contaminant source and aquifer geometry in a sandbox experiment with the restart ensemble kalman filter. J. Hydrol. 564, 1074–1084.

Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide, second ed. Oxford University Press, New York.

Evensen, G., 2009. Data Assimilation: The Ensemble Kalman Filter. Data Assimilation: The Ensemble Kalman Filter.

Ju, L., Zhang, J., Meng, L., et al. 2018. An adaptive Gaussian process-based iterative ensemble smoother for data assimilation. Adv. Water Resour., 115: 125-35.

Li, X., Chen, X., Hu, B.X., Navon, I.M., 2013a. Model reduction of a coupled numerical model using proper orthogonal decomposition. J. Hydrol. 507, 227–240.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Li, X., Hu, B. X., 2013b. Proper orthogonal decomposition reduced model for mass transport in heterogeneous media. Stochastic Environmental Research and Risk Assessment, 27(5): 1181-91.

Luo, Z., Li, H., Zhou, Y., et al. 2012. A reduced finite element formulation based on POD method for two-dimensional solute transport problems. Journal of Mathematical Analysis and Applications, 385(1): 371-83.

Luo, X., Bhakta, T., 2020. Automatic and adaptive localization for ensemble-based history matching. Journal of Petroleum Science and Engineering, 184: 106559.

Mo, S., Zabaras, N., Shi, X., et al. 2019. Deep Autoregressive Neural Networks for High-Dimensional Inverse Problems in Groundwater Contaminant Source Identification. Water Resour. Res., 55(5): 3856-81.

Pasetto, D., Guadagnini, A., Putti, M., 2011. 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. DOI:10.1016/j.advwatres.2011.07.003

Pasetto, D., Putti, M., Yeh, W.W.G., 2013. A reduced-order model for groundwater flow equation with random hydraulic conductivity: Application to Monte Carlo methods. Water Resour. Res., 49(6): 3215-3228. DOI:10.1002/wrcr.20136

Pasetto, D., Guadagnini, A., Putti, M., 2014. A reduced-order model for Monte Carlo simulations of stochastic groundwater flow. Computational Geosciences, 18(2): 157-169. DOI:10.1007/s10596-013-9389-4

Pasetto, D., Ferronato, M., Putti, M., 2017. A reduced order model-based

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

preconditioner for the efficient solution of transient diffusion equations. International Journal for Numerical Methods in Engineering, 109(8): 1159-1179. DOI:10.1002/nme.5320

Pinnau, R., 2008. Model Reduction via Proper Orthogonal Decomposition /Schilders, W. H. A., Van Der Vorst, H. A., Rommes, J. Model Order Reduction: Theory, Research Aspects and Applications. Berlin, Heidelberg; Springer Berlin Heidelberg. 95-109.

Razavi, S., Tolson, B.A., Burn, D.H., 2012. Review of surrogate modeling in water resources. Water Resour. Res. 48

Rizzo, C., de Barros, F., Perotto, S., Oldani, L., Guadagnini, A., 2018. Adaptive POD model reduction for solute transport in heterogeneous porous media. Computat. Geosci. 22, 297–308.

Stanko, Z.P., Boyce, S.E., Yeh, W.W.-G., 2016. Nonlinear model reduction of unconfined groundwater flow using pod and deim. Adv. Water Resour. 97, 130–143.

Tong, J., Hu, B. X., Yang, J., 2010. Using data assimilation method to calibrate a heterogeneous conductivity field conditioning on transient flow test data. Stochastic Environmental Research and Risk Assessment, 24(8): 1211-23.

Xia, C.-A., Luo, X., Hu, B.X., Riva, M., Guadagnini, A., 2021. Data assimilation with multiple types of observation boreholes via the ensemble Kalman filter embedded within stochastic moment equations. Hydrol. Earth Syst. Sci., 25(4): 1689-1709. DOI:10.5194/hess-25-1689-2021

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Xia, C.-A., Pasetto, D., Hu, B.X., Putti, M., Guadagnini, A., 2020. Integration of moment equations in a reduced-order modeling strategy for Monte Carlo simulations of groundwater flow. J. Hydrol., 590: 125257. DOI:https://doi.org/10.1016/j.jhydrol.2020.125257

Xia, C.-A., Guadagnini, A., Hu, B. X., Riva, M., Ackerer, P., 2019. Grid convergence for numerical solutions of stochastic moment equations of groundwater flow, Stoch. Environ. Res. Risk Assess., 33 (8-9), 1565-1579, https://doi.org/10.1007/s00477-019-01719-6.

Xia, C.-A., Hu, B.X., Tong, J., Guadagnini, A., 2018. Data Assimilation in Density-Dependent Subsurface Flows via Localized Iterative Ensemble Kalman Filter. Water Resour. Res., 54(9): 6259-6281. DOI:10.1029/2017wr022369

Xia, C.-A., Li, J., Riva, M., et al., 2024. Characterization of conductivity fields through iterative ensemble smoother and improved correlation-based adaptive localization. J. Hydrol., 634: 131054.

Xia, C.-A., Wang, H., Jian, W., et al., 2025. Reduced-order Monte Carlo simulation framework for groundwater flow in randomly heterogeneous composite transmissivity fields. J. Hydrol., 651: 132593.

Xu, T., Jaime Gomez-Hernandez, J., 2018. Simultaneous identification of a contaminant source and hydraulic conductivity via the restart normal-score ensemble Kalman filter. Adv. Water Resour., 112: 106-123. DOI:10.1016/j.advwatres.2017.12.011

Zhang, D., 2002. Stochastic Method for Flow in Porous Media – Coping with

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Uncertainties. Academic Press, Sand Diego, California.

Zhang J, Lin G, Li W, et al. An Iterative Local Updating Ensemble Smoother for Estimation and Uncertainty Assessment of Hydrologic Model Parameters With Multimodal Distributions. Water Resour Res, 2018, 54(3): 1716-33.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Figures

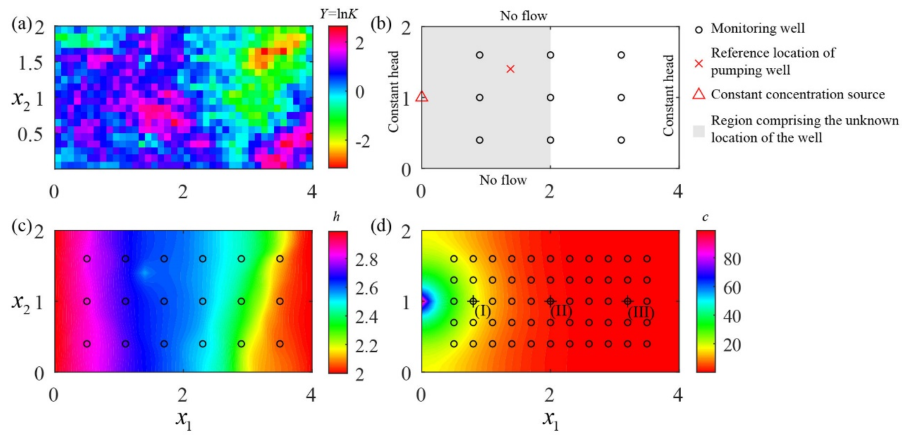

Fig. 1 (a) Reference field of Y = ln$K$; (b) boundary conditions for groundwater flow an solute transport together with spatial distribution of 9 monitoring wells and reference location for the pumping well (shaded gray area corresponds to the region comprising the unknown location of the well); (c) hydraulic head corresponding to the reference Y field; and (d) solute concentration corresponding the reference Y field at final time step, including three selected locations (i.e., I, II, and III) at which empirical probability density functions of solute concentration is computed and considered for illustration purposes. Circles in (b), (c) and (d) correspond to the location of the 9, 18 and 55 monitoring wells, respectively, employed in the study (see Section 3 and Table 1).

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

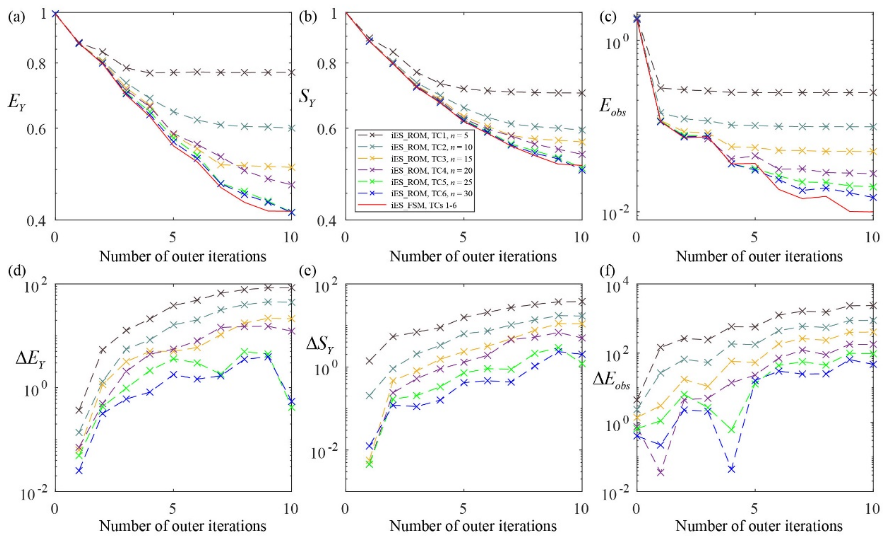

Fig. 2 Values of (a) E_Y, (b) S_Y, and (c) E_{obs} versus the number of outer iterations obtained through iES_ROM considering various dimensions of reduced-order model (with n = 5, 10, 15, 20, 25, and 30 for TCs 1-6, respectively) and iES_FSM (which provides identical results for TCs 1-6) for ensemble size N_{MC} = 10,000; corresponding percentage differences between the values of (d) E_Y (ΔE_Y), (e) S_Y (ΔS_Y), and (f) E_{obs} (ΔE_{obs}) evaluated through iES_ROM and iES_FSM.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

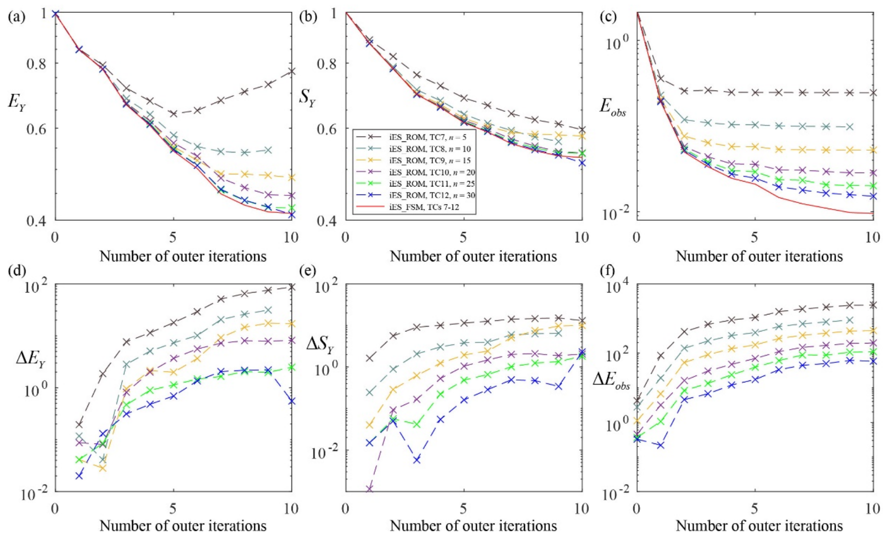

Fig. 3 Values of (a) E_Y, (b) S_Y, and (c) E_{obs} versus the number of outer iterations obtained through iES_ROM considering various dimensions of reduced-order model (with n = 5, 10, 15, 20, 25, and 30 for TCs 7-12, respectively) and iES_FSM (which provides identical results for TCs 7-12), when the pumping rate and location are previously known and for an ensemble size N_{MC} = 10,000; corresponding percentage differences between the values of (d) E_Y (ΔE_Y), (e) S_Y (ΔS_Y), and (f) E_{obs} (ΔE_{obs}) evaluated through iES_ROM and iES_FSM.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

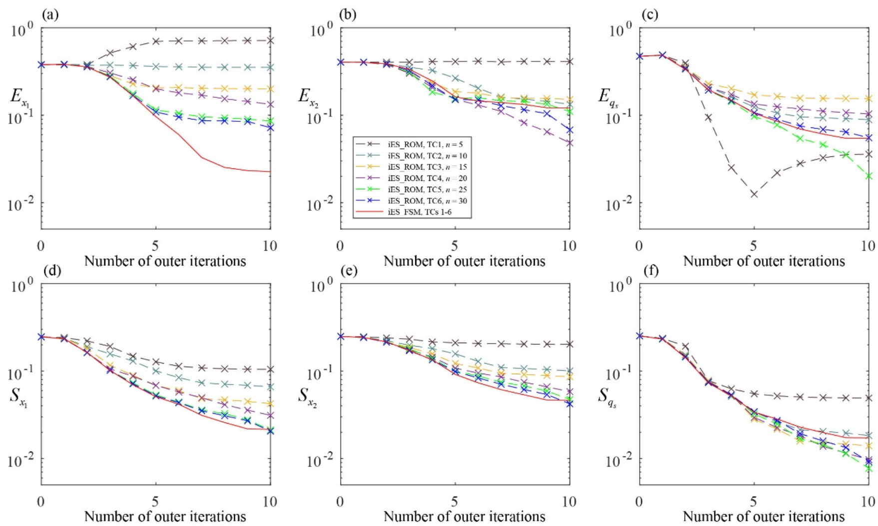

Fig. 4 Values of (a) E_{x_1}, (b) E_{x_2}, (c) E_{q_s}, (d) S_{x_1}, (e) S_{x_2}, and (f) S_{q_s} versus the number of outer iterations obtained through iES_ROM considering various dimensions of reduced-order model (with n = 5, 10, 15, 20, 25, and 30 for TCs 1-6, respectively) and iES_FSM (which provides identical results for TCs 1-6), when N_{MC} = 10,000.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

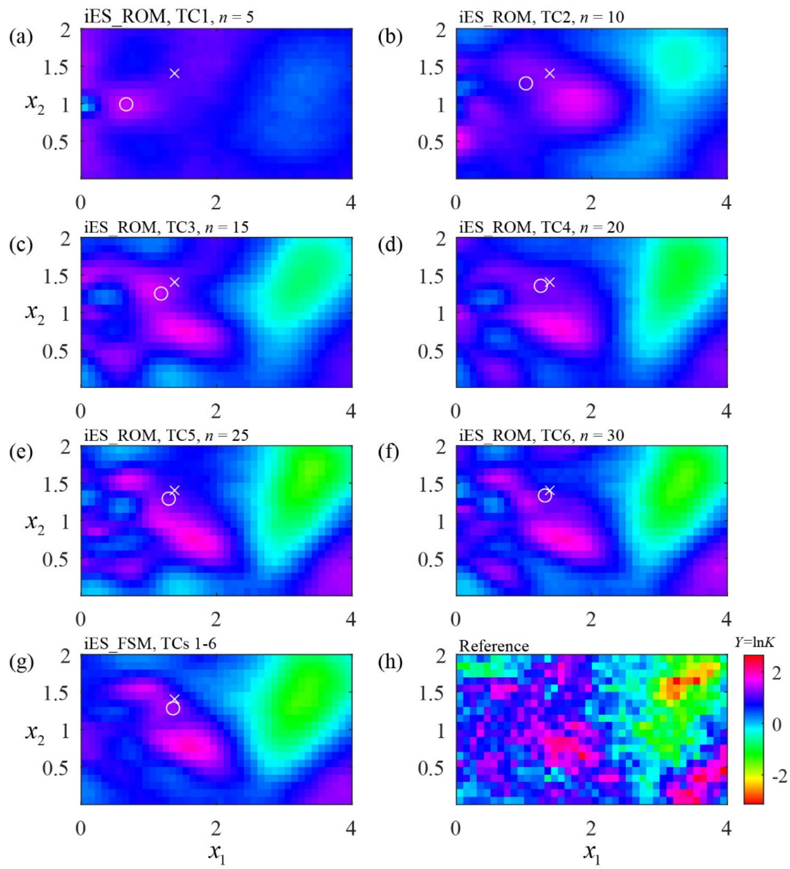

Fig. 5 Estimated (ensemble) mean Y fields at the final outer iteration through iES_ROM considering different n (equal to (a) 5, (b) 10, (c) 15, (d) 20, (e) 25, and (f) 30 for TCs 1-6, respectively) and (g) iES_FSM (which provides identical results for TCs 1-6) when N_{MC} = 10,000; (h) reference Y field.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

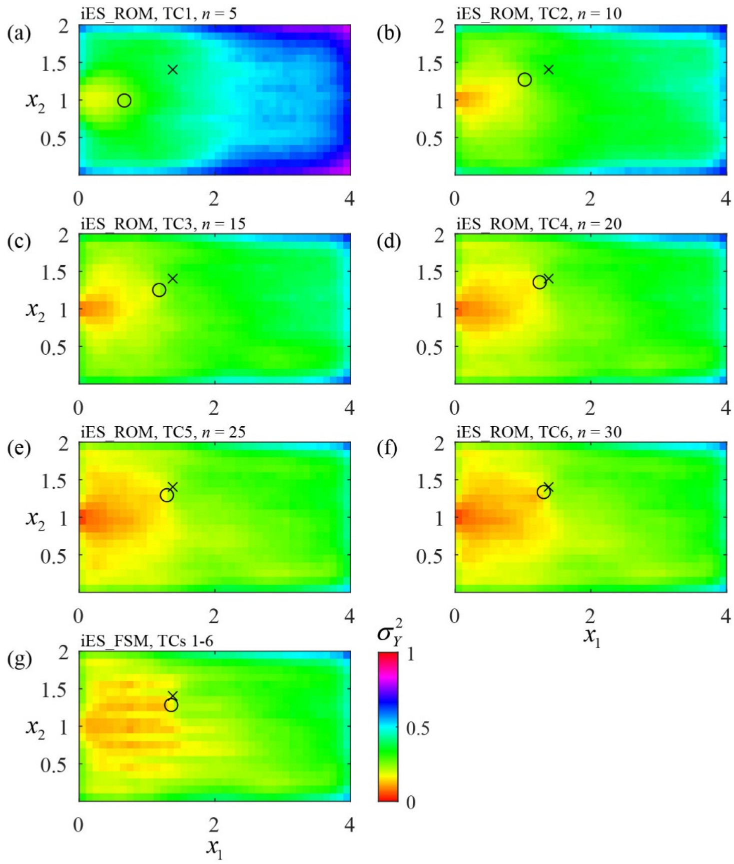

Fig. 6 Estimated (ensemble) Y variance fields at the final outer iteration through iES_ROM considering different n (equal to (a) 5, (b) 10, (c) 15, (d) 20, (e) 25, and (f) 30 for TCs 1-6, respectively) and (g) iES_FSM (which provides identical results for TCs 1-6), when N_{MC} = 10,000.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

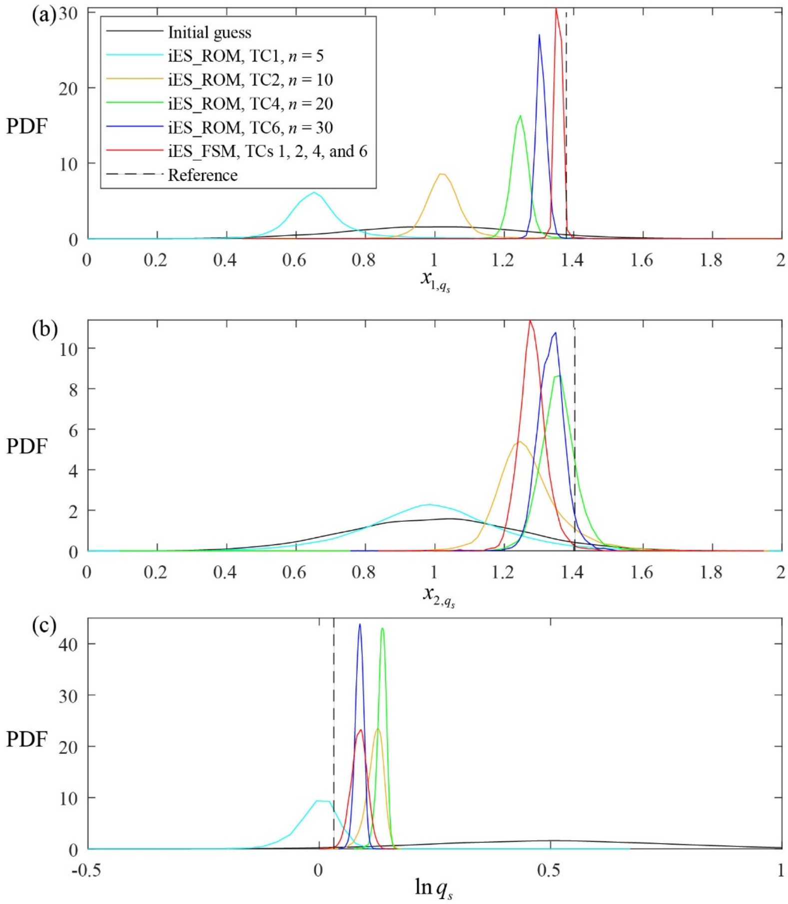

Fig. 7 Empirical PDFs of (a) x_{1,q_s}, (b) x_{2,q_s}, and (c) $\ln q_s$ at the final outer iteration through iES_ROM considering various values of n (equal to 5, 10, 20, and 30 for TCs 1, 2, 4, and 6, respectively) and iES_FSM (which provides identical results for TCs 1, 2, 4, and 6) when N_{MC} = 10,000; corresponding reference values are indicated by black dashed lines.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

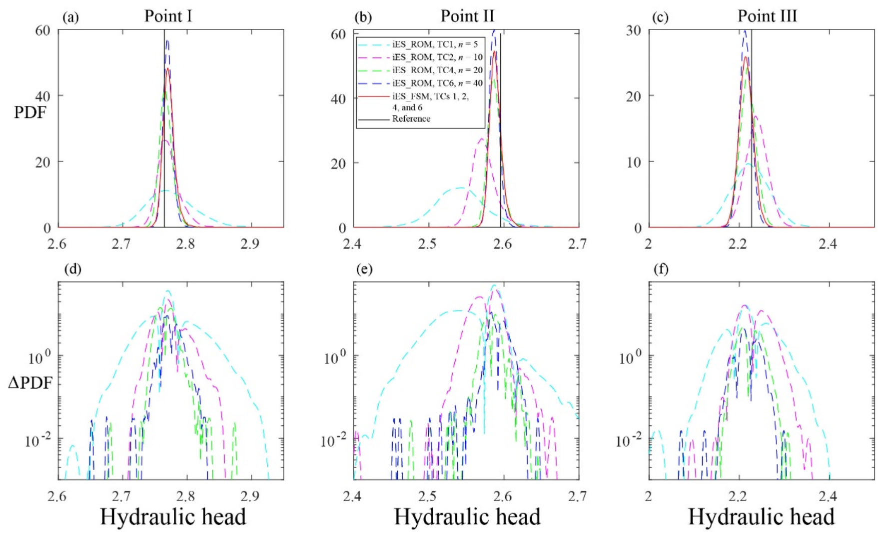

Fig. 8 Empirical PDFs of hydraulic head at points (a) I, (b) II, and (c) III (see Fig. 1) at the final outer iteration obtained through iES_ROM considering different values of n (equal to 5 (cyan dashed curve), 10 (magenta), 20 (green), and 30 (blue) for TCs 1, 2, 4, and 6, respectively) and iES_FSM (red solid curve; identical results for TCs 1, 2, 4, and 6) when N_{MC} = 10,000 (corresponding reference values are indicated by black vertical lines); logarithmic absolute difference between PDFs obtained through iES_ROM and iES_FSM at points (a) I, (b) II, and (c) III.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

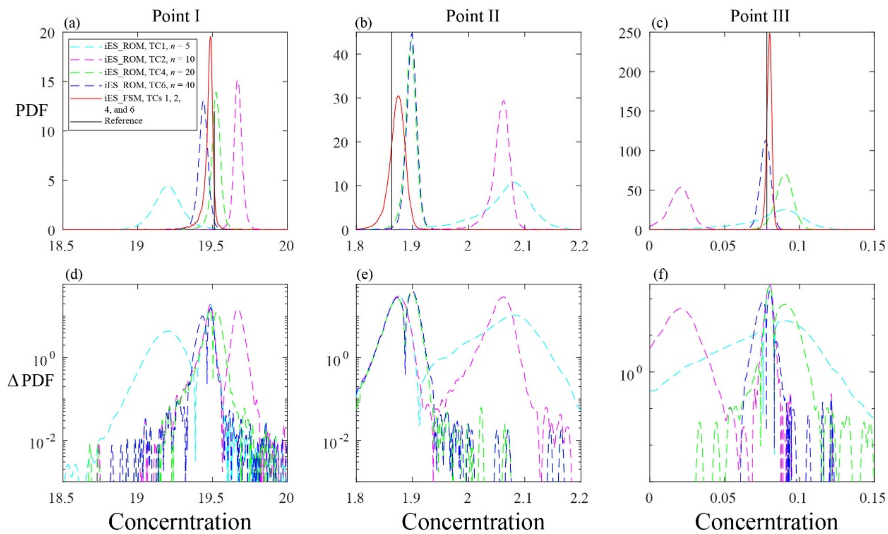

Fig. 9 Empirical PDFs of solute concentration at points (a) I, (b) II, and (c) III (see Fig. 1) at the final outer iteration obtained through iES_ROM considering various values of n (equal to 5 (cyan dashed curve), 10 (magenta), 20 (green), and 30 (blue) for TCs 1, 2, 4, and 6, respectively) and iES_FSM (red solid curve; results coincide for TCs 1, 2, 4, and 6) when N_{MC} = 10,000 (corresponding reference values are indicated by black lines); logarithmic absolute difference between PDFs obtained through iES_ROM and iES_FSM at points (a) I, (b) II, and (c) III.

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

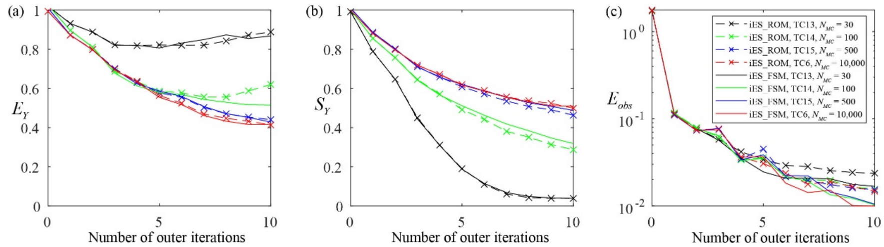

Fig. 10 Values of (a) E_Y, (b) S_Y, and (c) E_{obs} versus the number of outer iterations obtained through iES_ROM with n = 30 and iES_FSM considering various values of N_{MC} (equal to 30, 100, 500, and 10,000 for TCs 13-15 and 6, respectively).

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

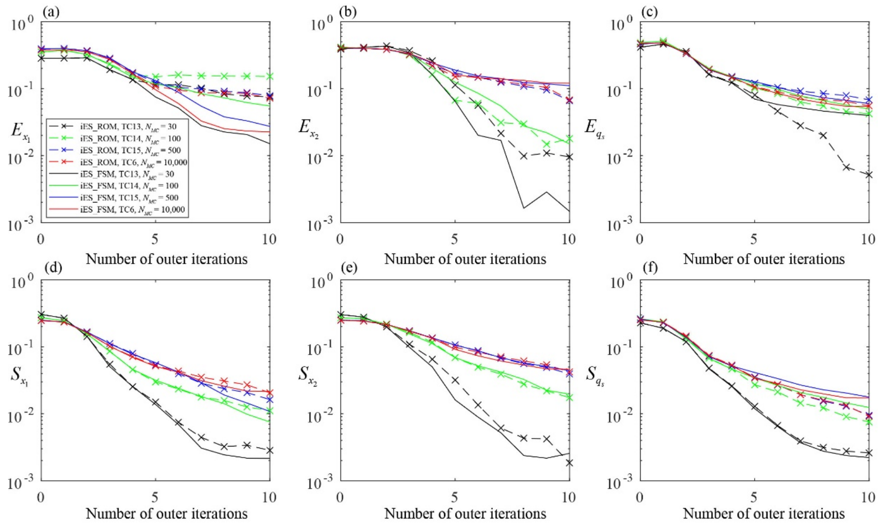

Fig. 11 Values of (a) E_{x_1}, (b) E_{x_2}, (c) E_{q_s}, (d) S_{x_1}, (e) S_{x_2}, and (f) S_{q_s} versus the number of outer iterations obtained through iES_ROM with n = 30 and iES_FSM considering various values of N_{MC} (equal to 30, 100, 500, and 10,000 for TCs 13-15 and 6, respectively).

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

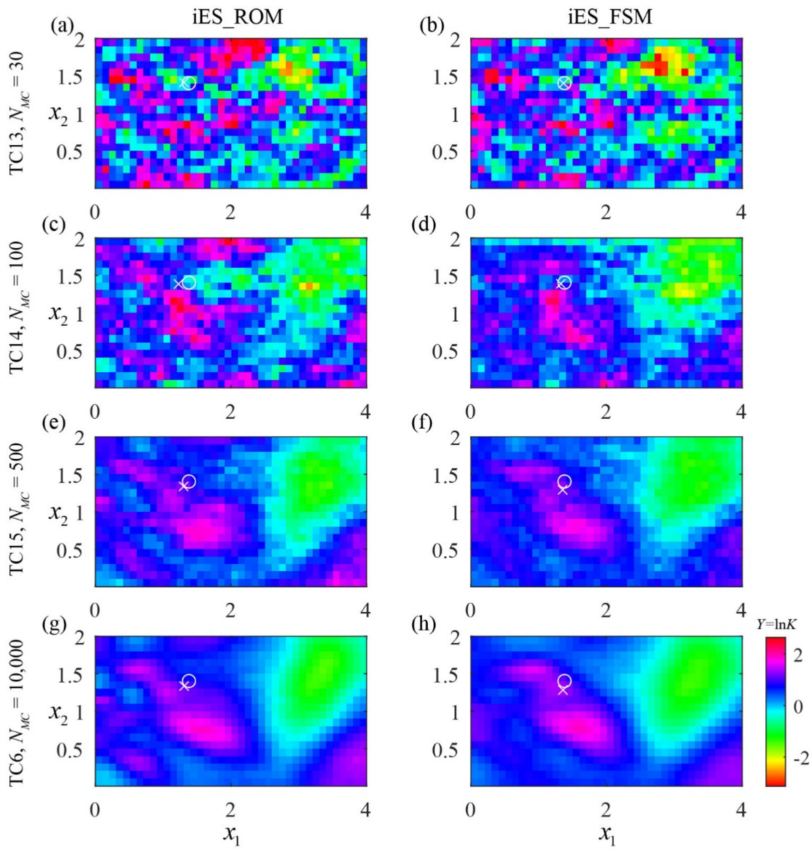

Fig. 12 Estimated (ensemble) mean Y fields at the final outer iteration obtained through iES_ROM with n = 30 (left column) and iES_FSM (right), considering N_{MC} = 30 (first row), 100 (second), 500 (third), and 10,000 (bottom) for TCs 13-15 and 6, respectively).

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

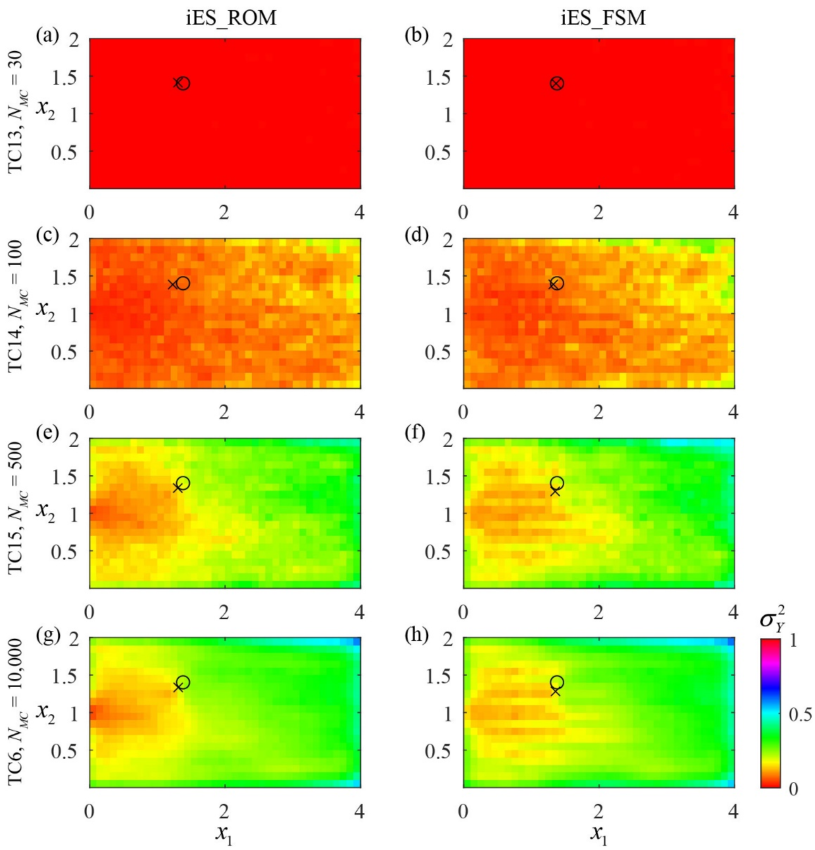

Fig. 13 Estimated (ensemble) Y variance fields at the final outer iteration obtained through iES_ROM with n = 30 (left column) and iES_FSM (right), considering N_{MC} = 30 (first row), 100 (second), 500 (third), and 10,000 (bottom) (corresponding to TCs 13-15 and 6, respectively).

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

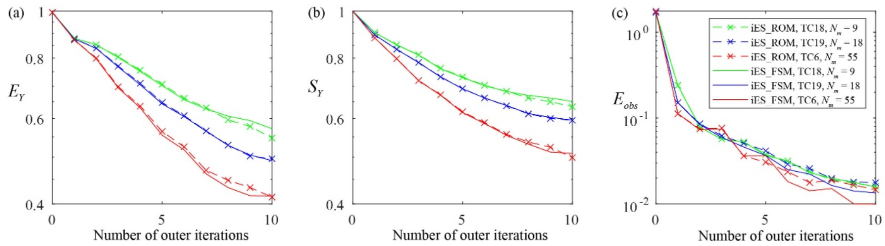

Fig. 14 Values of (a) E_Y, (b) S_Y, and (c) E_{obs} versus the number of outer iterations obtained through iES_ROM with n = 30 and iES_FSM considering N_m = 9, 18, and 55 (corresponding to TCs 18, 19, and 6, respectively)

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

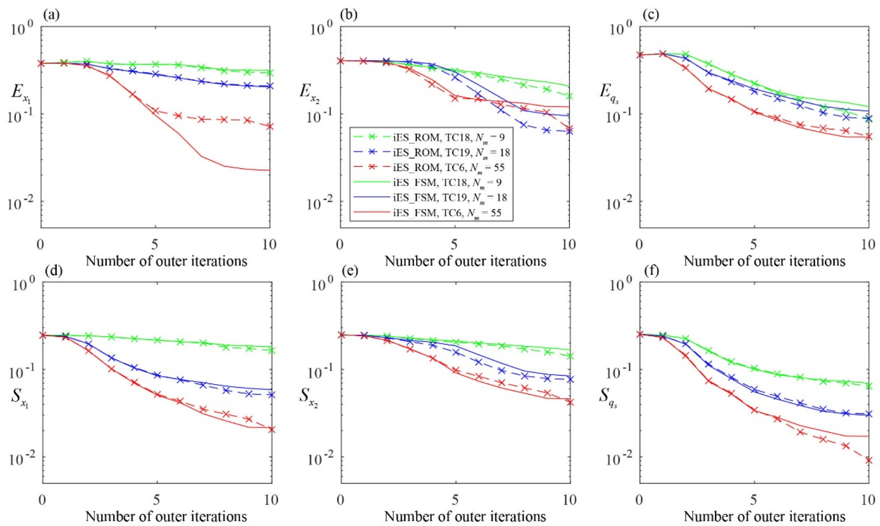

Fig. 15 Values of (a) E_{x_1}, (b) E_{x_2}, (c) E_{q_s}, (d) S_{x_1}, (e) S_{x_2}, and (f) S_{q_s} versus the number of outer iterations obtained through iES_ROM with n = 30 and iES_FSM, considering N_m = 9, 18, and 55 (corresponding to TCs 18, 19, and 6, respectively).

© Author(s) 2025. CC BY 4.0 License.

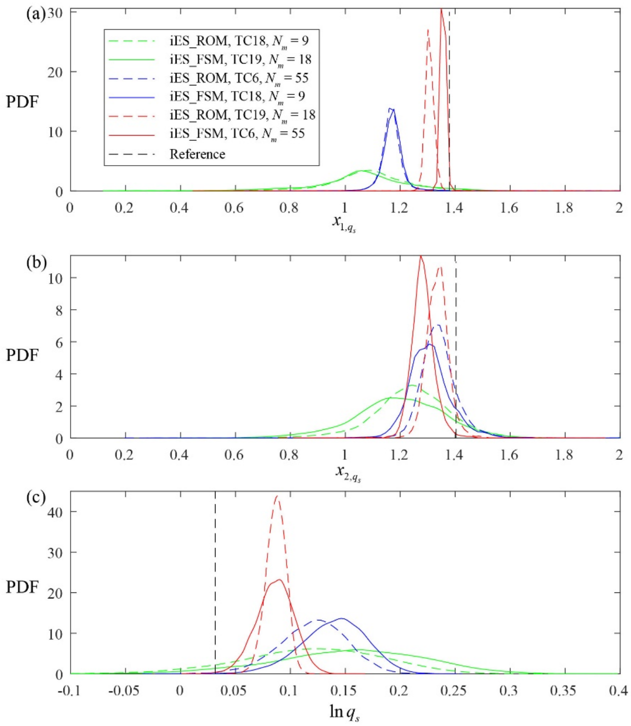

Fig. 16 Empirical PDFs of (a) x_{1,q_s}, (b) x_{2,q_s}, and (c) $\ln q_s$ at the final outer iteration through iES_ROM with n = 30 and iES_FSM (corresponding reference values indicated by black dashed lines) considering N_m = 9, 18, and 55 (corresponding to TCs 18, 19, and 6, respectively).

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Tables

Table 1 Overview of the key settings of the test cases (TCs) analyzed. All TCs are characterized by a zero mean and unit variance of the Y reference field; μ and σ_Y^2 denote the mean and variance of the initial ensemble of the Y fields, respectively.

Group A	Test Case	TC1, TC7	TC2, TC8	TC3, TC9	TC4, TC10	TC5, TC11	TC6, TC12
	n	5	10	15	20	25	30
	Known $q_s(\mathbf{x})$ or not	No, Yes	No, Yes	No, Yes	No, Yes	No, Yes	No, Yes
	Approach	iES_FSM and iES_ROM					
Group B	Test Case	TC13	TC14	TC15	TC6		
	N_{MC}	30	100	500	10,000		
	Approach	iES_FSM and iES_ROM					
Group C	Test Case	TC16	TC17	TC18	TC19	TC6	
	σ_{obs}	0.001	0.1	0.01	0.01	0.01	
	N_m	55	55	9	18	55	
	Approach	iES_FSM and iES_ROM					
Group D	Test Case	TC20	TC21	TC22	TC23	TC6	
	μ	-0.5	1.5	0.5	0.5	0.5	
	σ_Y^2	1.0	1.0	0.01	2.0	1.0	
	Approach	iES_FSM and iES_ROM					

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Group E	Test Case	TC24	TC25	TC26	TC27	TC28	TC6
	N_{sn}	30	100	300	500	1,000	10,000
	Approach	iES_ROM					

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Table 2 Values of E_Y, S_Y, E_{obs}, E_{x_1}, E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, and S_{q_s} at the end of the iteration procedure for TC16, TC6, and TC17 obtained through iES_ROM and iES_FSM.

	TC16	TC6	TC17	TC16	TC6	TC17	TC16	TC6	TC17
	E_Y			S_Y			E_{obs}		
iES_ROM	0.41	0.41	0.53	0.50	0.50	0.62	0.01	0.02	0.07
iES_FSM	0.41	0.42	0.52	0.51	0.51	0.60	0.01	0.01	0.07
	E_{x_1}			E_{x_2}			E_{q_s}		
iES_ROM	0.07	0.07	0.15	0.08	0.07	0.04	0.05	0.06	0.13
iES_FSM	0.02	0.02	0.06	0.13	0.12	0.09	0.05	0.05	0.11
	S_{x_1}			S_{x_2}			S_{q_s}		
iES_ROM	0.02	0.02	0.04	0.04	0.04	0.08	0.01	0.01	0.03
iES_FSM	0.02	0.02	0.04	0.05	0.05	0.07	0.02	0.02	0.03

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Table 3 Values of E_Y, S_Y, E_{obs}, E_{x_1}, E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, and S_{q_s} at the end of the iteration procedure for TC6, TC20, and TC21 obtained through iES_ROM and iES_FSM.

Test Case	TC20	TC6	TC21	TC20	TC6	TC21	TC20	TC6	TC21
	E_Y			S_Y			E_{obs}		
iES_ROM	0.51	0.41	0.50	0.60	0.50	0.52	0.02	0.02	0.03
iES_FSM	0.44	0.42	0.52	0.55	0.51	0.56	0.01	0.01	0.03
	E_{x_1}			E_{x_2}			E_{q_s}		
iES_ROM	0.03	0.07	0.12	0.10	0.07	0.10	0.06	0.06	0.08
iES_FSM	0.01	0.02	0.12	0.10	0.12	0.17	0.05	0.05	0.11
	S_{x_1}			S_{x_2}			S_{q_s}		
iES_ROM	0.05	0.02	0.03	0.08	0.04	0.06	0.03	0.01	0.01
iES_FSM	0.03	0.02	0.04	0.06	0.05	0.06	0.02	0.02	0.02

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Table 4 Values of E_Y, S_Y, E_{obs}, E_{x_1}, E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, and S_{q_s} at the end of the iteration procedure for TC6, TC22, and TC23 obtained through iES_ROM and iES_FSM.

Test Case	TC22	TC6	TC23	TC22	TC6	TC23	TC22	TC6	TC23
	E_Y			S_Y			E_{obs}		
iES_ROM	0.47	0.41	0.46	0.06	0.50	0.77	0.01	0.02	0.02
iES_FSM	0.43	0.42	0.47	0.06	0.51	0.80	0.01	0.01	0.01
	E_{x_1}			E_{x_2}			E_{q_s}		
iES_ROM	0.07	0.07	0.09	0.15	0.07	0.12	0.02	0.06	0.08
iES_FSM	0.02	0.02	0.05	0.22	0.12	0.17	0.02	0.05	0.08
	S_{x_1}			S_{x_2}			S_{q_s}		
iES_ROM	0.003	0.02	0.05	0.004	0.04	0.09	0.003	0.01	0.02
iES_FSM	0.002	0.02	0.04	0.003	0.05	0.08	0.003	0.02	0.03

https://doi.org/10.5194/egusphere-2025-5320
Preprint. Discussion started: 7 November 2025
© Author(s) 2025. CC BY 4.0 License.

Table 5 Percentage differences between the values of the selected metrics (i.e., E_Y, S_Y, E_{obs}, E_{x_1}, E_{x_2}, E_{q_s}, S_{x_1}, S_{x_2}, and S_{q_s}) at the end of the iteration procedure for TCs 24-28 obtained through iES_ROM (values corresponding to TC6 are taken as references).

Test Case	E_Y	S_Y	E_{obs}	E_{x_1}	S_{x_1}	E_{x_2}	S_{x_2}	E_{q_s}	S_{q_s}
TC24	11.88	6.34	21.60	44.17	58.50	25.04	34.75	32.29	55.20
TC25	10.16	3.66	6.76	27.83	35.54	13.24	9.58	25.01	37.58
TC26	7.40	1.97	13.00	22.10	16.89	35.54	11.26	13.06	15.09
TC27	4.58	0.14	2.66	11.42	0.17	22.93	1.22	17.74	3.37
TC28	0.50	0.21	4.18	17.19	1.33	31.33	5.86	14.03	0.07