Local and Global Error Models to improve uncertainty quantification

Laureline Josset[a,*], Ivan Lunati[a]

[a] *CRET, University of Lausanne, Switzerland*

Abstract

In groundwater applications, Monte Carlo methods are employed to model the uncertainty on geological parameters. However, their brute-force application becomes computationally prohibitive for highly detailed geological descriptions, complex physical processes, and a large number of realizations. The Distance Kernel Method (DKM) overcomes this issue by clustering the realizations in a multidimensional space based on the flow responses obtained by means of an approximate (computationally cheaper) model; then, the uncertainty is estimated from the exact responses that are computed only for one representative realization per cluster (the medoid). Usually, DKM is employed to decrease the size of the sample of realizations that are considered to estimate the uncertainty. We propose to use the information from the approximate responses for uncertainty quantification. The subset of exact solutions provided by DKM is then employed to construct an error model and correct the potential bias of the approximate model. Two error models are devised that both employ the difference between approximate and exact medoid solutions, but differ in the way medoid errors are interpolated to correct the whole set of realizations. The Local Error Model (LEM) rests upon the clustering defined by DKM and can be seen as a natural way to account for intra-cluster variability; the Global Error Model (GEM) employs a linear interpolation of all medoid errors regardless of the cluster to which the single realization belongs. These error models are evaluated for an idealized pollution problem in which the uncertainty of the breakthrough curve needs to be estimated. For this numerical test case, we demonstrate that the error models improve the uncertainty quantification provided by the DKM algorithm and are effective in correcting the bias of the estimate computed solely from the MsFV results. The framework presented here is not specific to the methods considered and can be applied to other combinations of approximate models and techniques to select a subset of realizations.

Keyword: multiscale finite volume, distance kernel method, stochastic simulation

*Corresponding author. CRET, University of Lausanne, Geopolis - UNIL Mouline, 1015 Lausanne, Switzerland. Tel.: +41-21-692-4418.

Email addresses: laureline.josset@unil.ch (Laureline Josset), ivan.lunati@unil.ch (Ivan Lunati)

1. Introduction

In groundwater applications one has to deal with an incomplete characterization of the aquifer: only sparse and uncertain measurements of the properties dictating the flow response is usually available. To account for this partial information, Monte Carlo methods are employed (Dagan 2002), which treat aquifer parameters, and in particularly the permeability (or equivalently the hydraulic conductivity), as stochastic variables. Several realizations of the permeability field, conditioned on the available data, are generated and the uncertainty is estimated from the variability of the responses obtained from different realizations. Despite the conceptual simplicity of this approach, the geostatistical representation of the uncertainty is rarely sufficient for realistically complex problems due to the large number of realizations required and the consequent prohibitive computational costs.

One possible strategy to overcome this issue is to employ approximate models that are less computationally expensive. Since in many applications large geological models are considered to describe the aquifer with high spatial resolution, one of the most effective techniques is to upscale the permeability on a coarser grid and solve reduced models. Several classical techniques exist at this end (Wen and Gómez-Hernández 1996; Renard and de Marsiliy 1997; Christie 1996; Durlofsky 2005); more modern multiscale approaches have been developed in the last decade that allow a better representation of the fine-scale details of the permeability field which are described by means of local numerical solution (Hou and Wu 1997; Arbogast 2002; Aarnes et al. 2005; Jenny et al. 2003).

The Multiscale Finite Volume (MsFV) method (Jenny et al. 2003) belongs to the latter group and has demonstrated great flexibility in modeling physically complex flows (Jenny et al. 2006; Lunati and Jenny 2006, 2007, 2008; Hajibeygi and Jenny 2009; Jenny and Lunati 2009; Künze and Lunati 2012). The accuracy of the MsFV method has been studied in a deterministic context and evaluated in terms of the ability to mimic the solution provided by the exact model in a single realization. This has fostered the development of several iterative strategies aimed at reducing these differences, which might be large in case of particularly challenging problems (Hajibeygi et al. 2008; Lunati et al. 2011; Zhou and Tchelepi 2012; Künze and Lunati 2012; Hajibeygi et al. 2012). In a stochastic context, however, a high level of accuracy might not be necessary because the goal is not to model each realization exactly, but simply to represent the variability of the ensemble of solutions (Chen and Durlofsky 2008; Chen et al. 2011; Aarnes and Efendiev 2008). As all methods that provide an approximate and relative inexpensive solution, the MsFV method is well suited to be applied in a stochastic context.

Another strategy to limit the computational cost of Monte Carlo approaches is to reduce the number of realizations for which the exact model is solved to estimate the uncertainty. Several methods exist to determine an optimal subset of realizations and coarsen the stochastic space. Some ranking methods classify the realizations based on static criteria such as geostatistical measures of connectivity or conductivity

(McLennan and Deutsch 2005). As they do not exploit information about the flow response, these methods are extremely efficient in terms of computational costs but have limited accuracy, which may result in a biased estimate of the uncertainty. Accuracy can be improved by using methods that sort the realizations based on a measure that depends on the flow response, such as in dynamic ranking methods (Ballin et al. 1992) or in the Distance Kernel Method (DKM) (Scheidt and Caers 2009a,b). While those approaches lead to much better results as they can be tailored to the question of interest, the problem remains of being able to inexpensively compute the dynamic measure.

In this paper, the MsFV method and the DKM are combined. However, rather than simply employing the MsFV method as approximate model to compute the dynamic measure in the DKM, the approximate MsFV solutions are used to obtain a first estimate of the uncertainty. The DKM selects a subset of realizations for which the exact model is solved; then, an error model to correct the potential bias of the MsFV estimate is constructed from the difference between the exact and the approximate solutions, which are available for the subset. Here, the ranking technique is used not solely to reduce the number of flow simulations, but rather to provide a representative subset of exact solutions to be compared to the approximate solutions. Note that whereas ranking techniques, or methods like DKM, make in general no direct use of the dynamic measure, in our approach this information is further exploited to construct an error model with negligible extra costs.

The paper is organized as follows: after a brief problem statement, we review the MsFV method and the DKM; then we present two error models that are devised by combining MsFV and DKM; finally, we present a thorough evaluation of the error models for a numerical test case that is representative of fluvio-glacial aquifers. The paper ends with some concluding remarks and perspectives for future development.

2. Problem statement

Here we consider the problem of predicting the breakthrough curve of a contaminant, which behaves as an ideal tracer (i.e., it does not alter the density and the viscosity of the fluid). The evolution of the contaminant concentration in the aquifer, c, is described by the following system of equations:

$$\nabla \cdot (K \nabla h) = 0 \tag{1}$$

$$\phi \frac{\partial c}{\partial t} + \nabla \cdot (c\boldsymbol{u} - \mathbf{D}\nabla c) = 0 \tag{2}$$

where

$$\boldsymbol{u} = -K\nabla h, \tag{3}$$

is the Darcy velocity; K the hydraulic conductivity (which is obtained dividing the permeability by the water viscosity); ϕ the porosity; and $\mathbf{D}$ the hydromechanical-dispersion tensor, which includes the effects

of molecular diffusion and dispersion. When appropriate boundary and initial conditions are assigned, the system above can be solved and the breakthrough curve at the location of interest can be computed as a function of time, $C(t)$.

The solution strongly depends on the structure of permeability and porosity fields (Lunati et al. 2003), which are usually not fully characterized on the basis of experimental observations. To model the uncertainty on these parameters, N_r realizations are generated, $\{K_i, \phi_i\}_{i=1,2,\ldots,N_r}$, which represent the variability of the properties due to the limited characterization of the aquifer. To evaluate the propagation of this uncertainty to the quantity of interest, flow and transport problems are solved in each realization and the breakthrough curve is computed, $C_i(t)$. (Here, initial and boundary conditions are treated as deterministic variables). The set of curves, $\{C_i(t)\}_{i=1,2,\ldots,N_r}$, obtained by these procedures, allows a characterization of the uncertainty on the breakthrough curve conditioned to the set of realizations that have been generated. In the following we are concerned with the problem of reducing the computational cost of these procedures, which can become prohibitive in presence of many geological realizations containing a large number of cells and involving complex physical processes.

3. Methodology

There are two natural strategies to overcome this issue: one is to use an approximate model that reduces the cost of computing a set of (approximate) curves $\{C_i^a(t)\}_{i=1,2,\ldots,N_r}$; the other is to reduce the dimensionality of the stochastic space and consider only a subset of $N_s < N_r$ realizations with breakthrough curves, $\{C_i(t)\}_{i=1,2,\ldots,N_s}$. Both strategies, however, might lead to biased predictions of the uncertainty.

The main idea of the present work is that the bias can be reduced by a combination of these two approaches. In the DKM, for instance, approximate models are used only to select the subset of realizations, $\{K_i, \phi_i\}_{i=1,2,\ldots,N_s}$, on the basis of their flow response. However, these approximate solutions can be used to estimate the variability neglected by the subset selection. On the other hand, the exact-model responses calculated for the selected realizations can be used to construct an error model and reduce the bias of the uncertainty estimated by the approximate model. In this paper we are precisely concerned with the problem of devising a methodology which allows an optimal exploitation of the information contained in the two sets of curves, that is $\{C_i^a(t)\}_{i=1,2,\ldots,N_r}$ and $\{C_i(t)\}_{i=1,2,\ldots,N_s}$.

3.1. The Multiscale Finite Volume (MsFV) method

The approximate model employed in this study is the MsFV method, which has been devised to efficiently solve the flow problem, Eq. (1), and deliver an approximate but fully conservative velocity field that can be used in the transport equation without introducing mass-balance errors (Jenny et al. 2003; Lunati and Jenny 2006). Although extensions of the MsFV method have been proposed in the past to solve the transport

problem (Lee et al. 2009; Künze and Lunati 2012), here the MsFV method is employed only to solve the flow problem, whereas the transport problem is solved exactly.

We use the operator formulations employed in Lunati and Lee (2009) to briefly present the MsFV method. First, we introduce the discrete form of Eq. (1)

$$\mathbf{A}\mathbf{h} = \mathbf{r}, \tag{4}$$

where $\mathbf{h}$ is the vector of the unknown hydraulic heads; $\mathbf{A}$ is the coefficient matrix, which depends on the hydraulic conductivity K; and $\mathbf{r}$ is the vector containing the information about the boundary conditions. In addition to the fine-scale grid introduced to define Eq. (4), the MsFV method employs two auxiliary coarse grids: a (primary) coarse grid and the corresponding dual (coarse) grid, which are represented in Fig. 1.

The main idea of the MsFV method is to approximate the hydraulic head by means of a set of interpolators, which are local numerical solutions computed on the cells of the dual grid, that is

$$\mathbf{h} \approx \mathbf{h}^{ms} = \mathbf{B}\mathbf{h}_n + \mathbf{C}\mathbf{r}, \tag{5}$$

where $\mathbf{B}$ is the basis-function operator, whose columns interpolate the hydraulic head, $\mathbf{h}_n$, at the node of the dual grid (which are at the centers of the coarse grid, see Fig. 1) to the fine-scale grid; $\mathbf{C}$ is the correction function operator, which accounts for the local effects of $\mathbf{r}$ and can be regarded as a source-term interpolator. In the MsFV method errors are introduced by the localization assumptions that are required to assign the boundary conditions of the local problems and compute basis and correction functions. Depending on flow conditions and on medium heterogeneity, localization might prevent a faithful description of long-correlation structures as channels or flow barriers (Lunati and Jenny 2004, 2007; Lunati et al. 2011).

The node hydraulic head, $\mathbf{h}_n$, is solution of the coarse equation

$$\mathbf{M}_{nn}\mathbf{h}_n = (\boldsymbol{\chi}\mathbf{A}\mathbf{B})\mathbf{h}_n = \boldsymbol{\chi}(\mathbf{I} - \mathbf{A}\mathbf{C})\mathbf{r}, \tag{6}$$

which is obtained by imposing the mass balance on the cells of the coarse grid (which serve as control volumes), that is by applying to $\mathbf{A}\mathbf{h}^{ms} = \mathbf{r}$ the summation operator, $\boldsymbol{\chi}$, which sums up all fine-cell values belonging to the same coarse cell and is the discrete analogous of control-volume integration. The computational advantage of the MsFV method stems from the fact that a large problem, Eq. (4), is split into a set of small local problems (which are solved to construct $\mathbf{B}$ and $\mathbf{C}$), and a coarse problem, Eq. (6), whose coefficient matrix, $\mathbf{M}_{nn} = \boldsymbol{\chi}\mathbf{A}\mathbf{B}$, is smaller than the original matrix, $\mathbf{A}$.

Once the approximate pressure solution, $\mathbf{h}^{ms}$, is obtained, a fine-scale conservative velocity field is constructed by solving a second set of local problems on the cells of the coarse grid and used in the transport equation. We refer to the existing literature for further details on the MsFV method (Lunati and Lee 2009 and references therein). Here, we simply remark that this framework offers great flexibility to implement

several adaptive strategies: the MsFV method can be seen as a numerical upscaling procedure, if the fine-scale velocity is not reconstructed and the transport is solved on the coarse grid (Lee et al. 2009; Künze and Lunati 2012); as an iterative linear solver, if a procedure is introduced to iteratively correct the boundary conditions of the localized problems (Hajibeygi et al. 2008; Lunati et al. 2011; Zhou and Tchelepi 2012); or as a downscaling method, if the original grid is taken as the coarse grid (Künze and Lunati 2012). Here, we use the MsFV method (with construction of a conservative velocity) as approximate model to compute a velocity field in each geostatistical realization; then, the MsFV approximate velocity is used in the transport equation, Eq. (2), to obtain a set of approximate breakthrough curves $\{C_i^{ms}(t)\}_{i=1,2,\ldots,N_r}$, which can be used to estimate the uncertainty.

3.2. Distance Kernel Methods (DKM)

DKM (Scheidt and Caers 2009a,b) is an alternative to traditional ranking techniques to select a subset of realizations that preserves the uncertainty spread of the sample. Dynamic ranking techniques (Ballin et al. 1992) sort realizations based on the responses of an approximate model and solve the exact model only for a subset of realizations that correspond to the desired quantiles. DKM, instead, employs the approximate information to quantify similarities between geostatistical models and selects a subset aiming at reproducing the same statistics as the full set of realizations. The first step is to compute a distance matrix $\mathbf{d}$ (a square matrix of size $N_r \times N_r$), which measures dissimilarity between realizations from the approximate flow responses. Here, the distance between two realizations, i and j, is defined as the l_2-distance between their breakthrough curves

$$d_{ij} = \sqrt{\sum_{t=1}^{n_t} [C_i^{ms}(t) - C_j^{ms}(t)]^2} \tag{7}$$

where $C_i^{ms}(t)$ is the curve obtained using MsFV as approximate model, and the sum is taken over all n_t discrete times at which the concentration is recorded (in our case the n_t time steps of the simulation). Eq. (7) naturally defines a multidimensional space, $\mathcal{S}$, where each realization is represented by a point and the distance between points is proportional to their dissimilarity in term of breakthrough response. It is natural to attempt to coarsen the space of uncertainty by grouping the realizations into N_s clusters based on their distances and assume that each cluster, Γ_k, can be represented by a representative realization (e.g., the medoid) weighted by the number of realizations in the cluster, N_{Γ_k}.

In DKM the clustering is not applied directly in the original multidimensional space, $\mathcal{S}$, but a kernel expansion is used to project the points onto a new space (the feature space $\mathcal{F}$) in the attempt to linearize the space of uncertainty. Although the expansion is associated with a kernel function of the form $\kappa[C_i^{ms}(t), C_j^{ms}(t)] = \langle \varphi[C_i^{ms}(t)], \varphi[C_j^{ms}(t)] \rangle$, where φ is the mapping function from $\mathcal{S}$ to $\mathcal{F}$, the distance matrix in the feature space, $\mathbf{d}^{\mathcal{F}}$, can be computed without an explicit definition of φ by using only the

scalar product computed by κ. Then the distance in the feature space is written as

$$d_{ij}^{\mathcal{F}} = \sqrt{K_{ii} + K_{jj} - 2K_{ij}} \tag{8}$$

where $\mathbf{K}$ is the kernel matrix associated to the kernel function. Among the many possible choices of the kernel matrix, we use a standard gaussian kernel of the form

$$K_{ij} = \exp\left\{\frac{-{d_{ij}}^2}{2\sigma^2}\right\} \tag{9}$$

where σ is the kernel width parameter.

Based on $\mathbf{d}^{\mathcal{F}}$, a k-medoid clustering algorithm (Hastie et al. 2009) is applied to find the many-to-one mapping, f, that assigns each curve, $C_i^{ms}(t)$, to a cluster (i.e., $f(i) = k$ if $C_i^{ms}(t) \in \Gamma_k$). The mapping corresponds to an optimization procedure, which finds

$$f = \arg\min_{f} \sum_{i,j:f(i)=f(j)} \mathbf{d}_{ij}^{\mathcal{F}} \tag{10}$$

and minimizes the average intra-cluster distances. In parallel to the definition of clusters, the algorithm identifies the medoids as the realizations that satisfies

$$i_k = \arg\min_{i:f(i)=k} \sum_{j:f(j)=k} \mathbf{d}_{ij}^{\mathcal{F}}. \tag{11}$$

The main advantage of k-medoids over k-means is that it does not require to explicitly compute points in the feature space and employs only the distance matrix in that space (Hastie et al. 2009). Moreover, k-medoids is not limited to Euclidean distances as k-means. This gives some freedom in defining the choice of the dissimilarity measure, which can be adapted to the question of interest.

The medoids define a subset of realizations, $\{K_{i_k}, \phi_{i_k}\}_{k=1,2,\ldots,N_s}$, for which the exact flow model is solved and a subset of exact curves $\{C_{i_k}(t)\}_{k=1,2,\ldots,N_s}$ is obtained. Classical DKM uses solely $\{C_i(t)\}_{i=1,2,\ldots,N_s}$ to compute experimental quantiles (Scheidt and Caers 2009a,b, 2010). This is done by assuming that all the realizations behave as the medoid realization, which leads to compute the experimental quantiles by weighting the medoid curves by the number of realizations in their cluster (or in other words, by considering a multiset of medoid curves, each having multiplicity equal to the number of cluster elements).

3.3. Error models

With the techniques described above, two sets of curves can be used to estimate the uncertainty of the predicted breakthrough curve that is $\{C_i^{ms}(t)\}_{i=1,2,\ldots,N_r}$ and $\{C_{i_k}(t)\}_{k=1,2,\ldots,N_s}$. In both cases, a sample of N_r realizations

$$\{C_i^*(t)\}_{i=1,2,\ldots,N_r}, \tag{12}$$

is used to compute experimental quantiles. If one choose to use only the approximate curves

$$\text{MsFV} \quad : \qquad C_i^*(t) = C_i^{ms}(t), \tag{13}$$

the MsFV uncertainty estimation is obtained. Employing the standard DKM is equivalent to choose

$$\text{DKM} \quad : \qquad C_i^*(t) = C_{i_k}(t), \quad \text{with} \quad k = f(i), \tag{14}$$

which construct a multiset where each medoid has multiplicity equal to the number of realizations in its cluster.

When the DKM is employed, information from approximate and exact responses is available and can be combined to improve uncertainty quantification at almost zero additional costs. On one hand, the information contained in the approximate curves can be used to estimate the intra-cluster variability, which is completely neglected by Eq. (14): the variability of cluster can be represented by the differences between each approximate curve and the approximate curve of its medoid, $C_i^{ms}(t) - C_{i_k}^{ms}(t)$. On the other hand, the exact curves of the medoids can be used to construct an error model aimed at reducing potential biases of the MsFV estimate: the difference between the exact and the approximate curves of the medoids, $C_{i_k}(t) - C_{i_k}^{ms}(t)$, can be used to correct all the curves in the cluster. These conceptually different approaches lead to exactly the same corrected curves

$$C_i^*(t) = C_{i_k}(t) + [C_i^{ms}(t) - C_{i_k}^{ms}(t)] = C_i^{ms}(t) + [C_{i_k}(t) - C_{i_k}^{ms}(t)] \tag{15}$$

with $k = f(i)$.

An error model of this form has been proposed in Scheidt et al. (2011) to estimate an upscaling error that is assumed to be the same for all realizations in the same cluster. In Scheidt et al. (2011), however, the corrected curves are used to generate realizations constrained to dynamic data. Notice that, if applied directly, Eq. 16 might lead to corrected curves that are unphysical and not constrained between zero and one. This is a severe limitation if the corrected curves are used to obtain an estimate of the uncertainty. To avoid this problem, the breakthrough curves are not corrected directly: first a logistic transformation is applied to all curves, $\hat{C}_i = \text{logit}^{-1}(C_i)$; then the transformed curves are corrected, $\hat{C}_i^*$; and finally, the corrected curves are transformed back via logit transformation, $C_i^* = \text{logit}(\hat{C}_i^*)$. This yields the Local Error Model (LEM)

$$\text{LEM:} \qquad C_i^*(t) = C_i^{lem}(t) = \text{logit}\left\{\hat{C}_i^{ms}(t) + [\hat{C}_{i_k}(t) - \hat{C}_{i_k}^{ms}(t)]\right\}, \tag{16}$$

which delivers corrected curves that lay between zero and one.

The error model above, which considers only intra-cluster information, can be readily extended by considering a set of linear combinations of corrected curves

$$C_i^*(t) = \sum_k^{N_s} \beta_{ik}\{C_i^{ms}(t) + [C_{i_k}(t) - C_{i_k}^{ms}(t)]\}, \tag{17}$$

where the weights, β_{ik}, might be chosen to enforce that the corrected curves have some desired characteristics (e.g., that they are constrained between zero and one, or that they are monotonic). Although the choice of the weighting function might be critical, here we chose a simple weighting function that depends exclusively on the distance in the feature space

$$\beta_{ik} = \frac{\exp(-d^{\mathcal{F}}_{ii_k})}{\sum_k^{N_s} \exp(-d^{\mathcal{F}}_{ii_k})}. \tag{18}$$

The underlying assumption is that realizations that are closer in the feature space have more similar errors. As for the LEM, to guarantee concentration values constrained between zero and one the logistic transformation used before applying the GEM and the corrected transformed curves are then transformed back via a logit transformation. This yields the Global Error Model (GEM)

$$\text{GEM:} \qquad C_i^*(t) = C_i^{gem}(t) = \text{logit}\left\{\hat{C}_i^{ms}(t) + \sum_k^{N_s} \beta_{ik}[\hat{C}_{i_k}(t) - \hat{C}_{i_k}^{ms}(t)]\right\}, \tag{19}$$

where it is assumed that $\sum_k \beta_{ik} = 1$, and observed that $C_i^{ms}(t)$ is independent of k.

Eq. 19 can be interpreted as an error model for the MsFV method. The exact curves computed for the N_s medoids are compared with the approximate curves of the medoids, and their difference is used to correct the approximate solution for each realizations i. Note that for an arbitrary weight, β_{ik}, all the medoid differences are used to correct each approximate curve. If $\beta_{ik} = \delta_{i,f(i)}$, the GEM reduces to the LEM and only intra-cluster information is used. If the constraint $\sum_k \beta_{ik} = 1$ is relaxed, the MsFV estimate of the uncertainty can be obtained by choosing $\beta_{ik} = 0$. A flowchart of the uncertainty analysis proposed here (which combines MsFV, DKM, and an error model) is presented in Fig. 2.

4. Numerical results

4.1. An idealized pollution problem

The methodology described above is applied to an idealized pollution problem in which the breakthrough curve of a contaminant has to be predicted. We consider a two-dimensional section of a confined aquifer of length 10.8 m and depth 5.1 m. The conductivity field, K, is inspired by the geology of a sedimentary aquifer, typical of braided river deposits. A vertical section acquired at the Herten site (Germany) (Bayer et al. 2011) is used as an input training image in the Direct Sampling method (MPDS) (Mariethoz et al. 2010) to perform multiple point geostatistical simulations and generate 1000 synthetic realizations. The 10 facies of the original data (Bayer et al. 2011) are reduced to 5 facies by grouping similar lithofacies. The porosity and of hydraulic conductivity values are reported in Fig. 3, together with the facies distribution of four realizations and the corresponding breakthrough curves.

No-flow conditions are applied at the upper and lower boundary of the domain, whereas two types of boundary conditions are considered for the left and right boundaries: prescribed incoming flux (BCF),

or prescribed hydraulic-head difference (BCH). The contaminant is released at the left boundary with normalized concentration $c = 1$, and the breakthrough curves are computed by averaging the concentration of the outcoming fluxes at the right boundary. In accordance with realistic natural gradient conditions simulations in which contaminant transport is dominated by advection (Péclet number Pe > 50) are run.

4.2. Application of the methodology

In this section, the methodology outlined in Fig. 2 is applied to the idealized pollution problem. Simulations with the exact model are performed on the full set of realizations and the variability of the responses, $\{C_i(t)\}_{i=1,2,\dots,N_r}$ (Fig. 4(a)), is taken has the reference uncertainty to evaluate the performance of the error models. Estimates provided by MsFV and DKM are also computed to illustrate the improvement achieved by LEM and GEM.

Experimental quantiles are calculated based on the approximate breakthrough curves, $\{C_i^{ms}(t)\}_{i=1,2,\dots,N_r}$ (Fig. 4(b)), and provide the MsFV estimate of the uncertainty. Then, a distance matrix is constructed using MsFV curves and DKM is applied to identified N_{Γ_k} clusters and select a subset of realizations. The number of clusters should be sufficient to capture the error and estimate the desired quantiles, but not too large in order to limit the computational costs. Although a procedure could be devised to identify an optimal number, here we simply set $N_{\Gamma_k} = 20$, which corresponds to a coarsening factor of 50 for the uncertainty space and allows computing the 10^{th} and 90^{th} percentiles (P10 and P90, respectively) by the DKM. The identification of the subset is performed in the feature space using a Gaussian-kernel expansion. After a sensitivity analysis, the width parameter is set equal to the standard deviation of the distance matrix, which 0.55 and 0.98 for BCF and BCH, respectively. The clustering is performed only on the base of the kernel matrix and does not require constructing the feature space explicitly. The k-medoids algorithm is used to identify N_{Γ_k} medoids for which the exact responses are computed, $\{C_{i_k}(t)\}_{k=1,2,\dots,N_{\Gamma_k}}$.

A two-dimensional representation of the clustering in the feature space is shown in Fig. 5. The realizations seem continuously distributed rather than arranged in well separated clusters. Although this might be partially due to the two-dimensional visualization of the feature space, the fact that clusters are not well defined is confirmed by the instability of the clustering algorithm: different initializations of the algorithm (which require an initial guess on the N_{Γ_k} medoids) lead to different cluster repartitions and different uncertainty predictions, independently of the kernel width choice. A set of exact breakthrough curves obtained for one of the cluster repartitions, $\{C_{i_k}(t)\}_{k=1,2,\dots,N_{\Gamma_k}}$, is shown in Fig. 4(c).

The approximate curves for the entire set of realizations and for the medoid exact response are then used to construct the error model. Here our approach differs from the standard DKM, which estimates the quantiles based exclusively on the subset of exact curves and does not make any direct use of the set of approximate curves. In contrast, we use the differences between the approximate and exact medoid responses to correct the entire set of approximate curves, which is then used to estimate the quantiles.

In the LEM the responses are corrected using only local (intracluster) information and the set of curves $\{C_i^{lem}(t)\}_{i=1,2,\ldots,N_r}$ (Eq. (16), Fig. 4(d)) is used to compute the quantile. In the GEM the responses are corrected globally, regardless to the cluster to which they belong, and the set of curves $\{C_i^{gem}(t)\}_{i=1,2,\ldots,N_r}$ (Eq. (19), Fig. 4(e)) is obtained. Notice that few outliers are not effectively corrected due to the limited coverage of the extreme regions by the set of medoids. As it will be seen in the next section, this few outliers do not sensitively affect the estimate of P10, P50, and P90. However, in cases where uncertainty on extremes needs to be quantified, a different strategy has to be used to identify the subset of realizations used to construct the error models and extreme regions have to be more densely sampled. Note that due to the non-clear repartitions of the realizations into well defined clusters, this global model is more consistent with the data and it is expected to lead to more stable uncertainty estimations in terms of dependency on the initial medoids guess.

4.3. Comparison of quantile-curve estimates

In general, the characterization of uncertainty is done on the basis of a limited number of experimental quantiles; here we consider the 10^{th}, 50^{th} and 90^{th} percentiles (P10, P50, and P90, respectively). Figs. 6 and 7 compare the three quantile curves obtained with the four models (MsFV, DKM, LEM and GEM) with the reference quantile curves for both sets of boundary conditions. Notice that due to the instability of the DKM algorithm, which depends on the initial guess on the medoids, very different quantile curves can be obtained with DKM, LEM, and GEM. Here we present the comparison for an initialization with yield an average performance, whereas the variability in model response due to the stability of DKM is investigated in the next section.

For BCF, MsFV provides a good measure of the statistical variability but tends to slightly underestimate contaminant concentration of about 4.5% at early times (note that the concentration will be overestimated at later time due to the constraint that the approximate MsFV solution is conservative and therefore the mean arrival time of the contaminant must be exact with this type of boundary conditions). DKM leads to curves that are less smooth due to the reduction of statistical space, which deteriorates the estimates of quantile curves; the average maximum fluctuations are of the order of 3%. LEM provides smoother curves than DKM, whereas GEM gives an excellent estimation of the uncertainty (average maximum fluctuations LEM and GEM are of 1.8% and 1.5% respectively). MsFV bias is effectively corrected and the uncertainty is correctly represented by the MsFV approximate curves.

For BCH, MsFV quantile curves are in good agreement with the reference (maximum difference between the curves is of 5.2%). DKM estimate is less good and the 20 exact medoid responses provide a worst uncertainty estimate than the set of approximate responses (fluctuations of 6.1%). This shows that in some case DKM can lead to a deteriorated prediction of approximate solutions on which it is based. LEM also smooths the DKM estimation for this set of boundary conditions, but P10 and P90 remain underestimate

(averaged maximum fluctuations of 4.3%); GEM leads again to an excellent estimate (3.3%).

4.4. Quantification of the quality of the estimate and stability

To illustrate the dependence of clustering on the initialization, DKM, LEM, and GEM are applied 500 times time with a different initial guess of the medoid set (seed). The overall quality of the different models is evaluated by considering the l_2-norm of the quantile error

$$l_2 : \qquad \sqrt{\sum_t (P_T(t) - P_E(t))^2}, \tag{20}$$

where $P_T(t)$ is the reference quantile curve and $P_E(t)$ is the estimated quantile.

Figs. 8 and 9 shows the errors for the two set of boundary conditions and for the 500 seeds. For each quantiles, the mean error of each method is represented by a bar plot, whereas the error bars represent the 80% confidence interval (i.e., the interval in which one finds 80% of the 500 results obtained with different seeds). These plots clearly show that the DKM error can be much larger than what observed in Figs. 6 and 7, which correspond to an initialization leading to an error close to the mean of the results from the 500 initializations.

For BCF, DKM performs, in average, better than MsFV for P50 and P90. MsFV responses provide an accurate selection of representative realizations, but yield a relatively poor estimate of the uncertainty due to the systematic underestimation of the concentration (see Fig. 6). However, DKM shows a large variability depending on the initialization of the clustering algorithm and for some seed can lead to larger errors than MsFV (1.7 times higher for P10 in 10% of the cases). LEM and GEM result in a much better estimate and lead to a considerable reduction of the dependency on the initialization of the algorithm. GEM performs better than LEM on both aspects (although for P90 GEM shows a slightly larger seed dependency).

For BCH, the MsFV estimate yields a sensibly lower error than the one obtained by DKM, and this despite the fact that information from 20 exact simulations is used in DKM. This is likely due to the large instability of the clustering algorithm that can lead to very unreliable estimates. This example clearly demonstrate how dangerous could be to rely only on medoid information, thus on an extremely small stochastic space, for estimate P10 and P90. The error models can correct this problem and lead to a better estimate than MsFV for GEM. For P90, one can observe a dramatic reduction of the seed dependency with respect to DKM, whose upper bound of the 80% confidence interval lays at 0.57; LEM reduces this to 0.39 and GEM to 0.16.

In conclusion, MsFV provides a good estimate of the statistical variability but tend to present some systematic bias. DKM provides a good subset of representative realizations, but is strongly affected by the reduction of statistics. Both error models improve substantially the quantification of uncertainty by combining the whole available information. They both lead to a reduction of dependency on the algorithm seed; and GEM provides an excellent and much more stable estimate in both situations.

4.5. Cumulative distribution function at a given time step

Finally, we consider the estimated Cumulative Distribution Function (CDF) at two time steps: $t = 70$ for BCF (Fig. 10), and $t = 14$ (Fig. 11) for BCH, respectively. The CDFs in Figs. 10 and 11 refer to a single initialization (seed) of DKM, which has been chosen to be representative of the average result. Depending on the cluster initialization, however, the quality of the DKM results would be different.

For fixed-flux boundary conditions (BCF), one can observe a systematic shift of the MsFV CDF towards smaller concentrations; whereas for fixed-head boundary conditions (BCH), the MsFV CDF is close to the reference. Depending on the percentile, the error of the DKM estimate could be as high as 5% of concentration for BCF and 12% for BCH. The DKM CDF exhibits a staircase behavior, which is the result of the clustering and the subsequent reduction of the number of realizations used to compute the CDF: the DKM estimate employs only the N_{Γ_k} medoid curves and neglects intra-cluster variability. This problem can be overcome by using the LEM or the GEM which construct a sample containing the same number of realizations as the original set, $\{C_i^{lem}(t)\}_{i=1,2,\ldots,Nr}$. As a consequence a smooth CDF is obtained and the error is reduced. GEM provides an excellent estimate of the CDF for BCF and does just as well as MsFV for BCH.

5. Conclusions

The DKM is applied to estimate uncertainty at lower computational costs than a brute-force Monte Carlo approach. The method relies on an approximate model to select a subset of representative realizations for which the exact model is solved; then, the uncertainty is estimated only on the basis of the exact-response subset with no additional use of the approximate solutions. This approach neglects intra-cluster variability, leads to a dimensional reduction of the statistical space, and provides uncertainty estimates with a lower resolution than allowed by the original set of realizations. For our numerical test case, the DKM is not stable with respect to the initialization (seed) of the clustering algorithm and that this can lead to inaccurate predictions: in most critical cases, the DKM can even deteriorate the uncertainty estimate provided solely by the approximate solutions. On the other hand, however, using only the approximate responses obtained with the MsFV method can lead to biased estimates of the uncertainty due to the localization assumptions, which reduce the accuracy of the solution in presence of long structures spanning several coarse cells. If this is an issue in a deterministic context (where iterative schemes are usually required to achieve the desired accuracy), in a stochastic framework this is a minor problem, which can be solved by means of an error model.

Two error models are devised that aim at exploiting the whole available information and combine the MsFV approximate responses with the exact responses obtained for medoids selected by the DKM. Both models employs the difference between approximate and exact solutions for the medoid realizations, but

differ in the way this discrepancy is interpolated to correct each realization. The LEM applies the same correction to all realizations belonging the same cluster and can be seen as a natural way to model the intra-cluster variability of the responses; the GEM corrects each realization by a linear interpolation of all medoid errors (weighted by a function of the distance in the feature space) regardless to the cluster to which it belongs. Both models improve the DKM estimate and reduce the dependency on the initialization of the clustering. The GEM leads to excellent uncertainty estimates and performs systematically better than the LEM; this is likely due to the fact that a global error model (which does not rely only on intra-cluster information) is more consistent with the data considered in this study, which are not separated in clearly defined clusters.

The framework presented here is not specific to the methods considered (namely MsFV and DKM) but can be applied to other combinations of approximate models and techniques to select a subset of realizations. For instance, it can be used in a multiphysics context where the approximate model employs a simplified physical description and an error model is developed to predict a more complicated physical process (e.g., single phase vs. multiphase flow problems).

Some of the steps can be extended and generalized to ameliorate the reliability of the error model for challenging test cases. In particular four main improvements can be suggested: the selection of the representative realizations can be modified to obtain a larger number of realizations in regions of interest rather than uniformly covering the entire feature space; the subset of representative realizations could be iteratively enlarged until a number of realizations is selected that allows the required level of accuracy (note that this would require an a-posteriori estimate of the accuracy to define the stopping criterion); the weights used in the global error model can be obtained from the solution of an optimization problem, which can be tailored to guarantee that the corrected responses satisfy certain physical constraints (this entails a more profound re-thinking of all steps to determine the ideal subset); finally, Functional Data Analysis (FDA) can be used to keep an explicit time dependence and work with breakthrough curves in a functional space rather than with points in a feature space.

Acknowledgments

Many thanks are also due to Rouven Künze for his assistance with the flow simulations, and Guillaume Pirot and Philippe Renard for providing the realizations of the hydraulic conductivity. This project is supported by the Swiss National Science Foundation as a part of the ENSEMBLE project (Sinergia Grant No. CRSI22-132249/1). The authors thank Céline Scheidt and and Jef Caers for sharing their DKM code and many useful discussions. Ivan Lunati is Swiss National Science Foundation (SNSF) Professor at the University of Lausanne (SNSF grant number PP00P2-123419/1).

References

Aarnes, J. and Efendiev, Y. (2008). "Mixed multiscale finite element methods for stochastic porous media flows." *SIAM Journal on Scientific Computing*, 30(5), 2319–2339.

Aarnes, J., Kippe, V., and Lie, K. (2005). "Mixed multiscale finite elements and streamline methods for reservoir simulation of large geomodel." *Adv. Water Res.*, 28, 257–271.

Arbogast, T. (2002). "Implementation of a locally conservative numerical subgrid upscaling scheme for two phase darcy flow." *Comput. Geosci.*, 6, 453–481.

Ballin, P., Journel, A., and Aziz, K. (1992). "Prediction of uncertainty in reservoir performance forecast." *Journal of Canadian Petroleum Technology*, 31(4).

Bayer, P., Huggenberger, P., Renard, P., and Comunian, A. (2011). "Three-dimensional high resolution fluvio-glacial aquifer analog: Part 1: Field study." *Journal of Hydrology*, 405(1), 1–9.

Borg, I. and Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*. Springer.

Chen, Y. and Durlofsky, L. (2008). "Ensemble-level upscaling for efficient estimation of fine-scale production statistics." *SPE Journal*, 13(4), 400–411.

Chen, Y., Park, K., and Durlofsky, L. J. (2011). "Statistical assignment of upscaled flow functions for an ensemble of geological models." *Computational Geosciences*, 15(1), 35–51.

Christie, M. (1996). "Upscaling for reservoir simulation." *Journal of Petroleum Technology*, 48(11), 1004–1010.

Cox, M. and Cox, T. (2008). "Multidimensional scaling." *Handbook of data visualization*, 315–347.

Dagan, G. (2002). "An overview of stochastic modeling of groundwater flow and transport: From theory to applications." *Eos, Transactions American Geophysical Union*, 83(53), 621.

Durlofsky, L. (2005). "Upscaling and gridding of fine scale geological models for flow simulation." *8th International Forum on Reservoir Simulation Iles Borromees, Stresa, Italy*. 20–24.

Hajibeygi, H., Bonfigli, G., Hesse, M., and Jenny, P. (2008). "Iterative multiscale finite-volume method." *J. Comp. Phys*, 227(19), 8604–8621.

Hajibeygi, H. and Jenny, P. (2009). "Multiscale finite-volume method for parabolic problems arising from compressible flow in porous media." *J. Comp. Phys*, 228, 5129–5147.

Hajibeygi, H., Lee, S. H., and Lunati, I. (2012). "Accurate and efficient simulation of multiphase flow in a heterogeneous reservoir by using error estimate and control in the multiscale finite volume method." *SPE J.*, SPE–141954–PA. In press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). "The elements of statistical learnin.

Hou, T. Y. and Wu, X. H. (1997). "A multiscale finite element method for elliptic problems in composite materials and porous media." *J. Comp. Phys*, 134(1), 169–189.

Jenny, P., Lee, S. H., and Tchelepi, H. (2003). "Multi-scale finite-volume method for elliptic problems in subsurface flow simulation." *J. Comp. Phys*, 187(1), 47–67.

Jenny, P., Lee, S. H., and Tchelepi, H. (2006). "Adaptive fully implicit multi-scale finite-volume method for multi-phase flow and transport in heterogeneous porous media." *J. Comp. Phys*, 217, 627–641.

Jenny, P. and Lunati, I. (2009). "Modeling complex wells with the multi-scale finite-volume method." *J. Comp. Phys*, 228(3), 687–702.

Künze, R. and Lunati, I. (2012). "An adaptive multiscale method for density-driven instabilities." *Journal of Computational Physics*.

Lee, S., Zhou, H., and Tchelepi, H. (2009). "Adaptive multiscale finite-volume method for nonlinear multiphase transport in heterogeneous formations." *Journal of Computational Physics*, 228(24), 9036–9058.

Lunati, I. and Jenny, P. (2004). "Multi-scale finite-volume method for flow in highly heterogeneous porous media with shale layers." *European Conference of Mathematics of Oil Recovery IX, Cannes, France*.

Lunati, I. and Jenny, P. (2006). "Multi-scale finite-volume method for compressible flow in porous media." *J. Comp. Phys*, 216(2), 616–636.

Lunati, I. and Jenny, P. (2007). "Treating highly anisotropic subsurface flow with the multiscale finite-volume method." *Multiscale Model. Simul.*, 6(1), 308–318.

Lunati, I. and Jenny, P. (2008). "Multiscale finite-volume method for density-driven flow in porous media." *Comput. Geosci.*, 12(3), 337–350.

Lunati, I., Kinzelbach, W., and Sørensen, I. (2003). "Effects of pore volume-transmissivity correlation on transport phenomena." *J. Cont. Hydr.*, 67(1-4), 195-217.

Lunati, I. and Lee, S. (2009). "An operator formulation of the multiscale finite-volume method with correction function." *Multiscale Model. Simul.*, 8(1), 96–109.

Lunati, I., Tyagi, M., and Lee, S. (2011). "An iterative multiscale finite-volume algorithm converging to the exact solution." *J. Comp. Phys*, 230(5), 1849–1864.

Mariethoz, G., Renard, P., and Straubhaar, J. (2010). "The direct sampling method to perform multiple-point geostatistical simulations." *Water Resources Research*, 46(11), W11536.

McLennan, J. and Deutsch, C. (2005). "Ranking geostatistical realizations by measures of connectivity." *SPE International Thermal Operations and Heavy Oil Symposium, 98168, Alberta, Canada.*

Renard, P. and de Marsiliy, G. (1997). "Calculating equivalent permeability: a review." 20(5-6), 253–278.

Scheidt, C. and Caers, J. (2009a). "Representing spatial uncertainty using distances and kernels." *Mathematical Geosciences*, 41(4), 397–419.

Scheidt, C. and Caers, J. (2009b). "Uncertainty quantification in reservoir performance using distances and kernel methods–application to a west africa deepwater turbidite reservoir." *SPE Journal*, 14(4), 680–692.

Scheidt, C. and Caers, J. (2010). "Bootstrap confidence intervals for reservoir model selection techniques." *Computational Geosciences*, 14, 369–382.

Scheidt, C., Caers, J., Chen, Y., and Durlofsky, L. J. (2011). "A multi-resolution workflow to generate high-resolution models constrained to dynamic data." *Computational Geosciences*, 15(3), 545–563.

Wen, X. and Gómez-Hernández, J. (1996). "Upscaling hydraulic conductivities in heterogeneous media: An overview." *J. Hydr.*, 183(1-2), R9–R32.

Zhou, H. and Tchelepi, H. (2012). "Two-stage algebraic multiscale linear solver for highly heterogeneous reservoir models." *SPE Journal*, 17(2), 523–539.

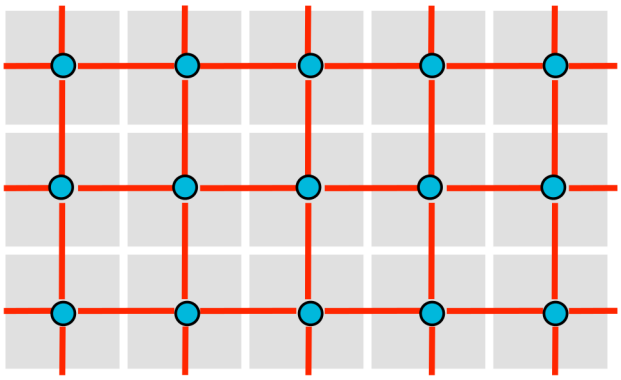


Figure 1: Representation of the auxiliary coarse grids used in the MsFV method. The dual (coarse) grid (red lines) is used to construct a set of local interpolators, which are local numerical solutions, whereas the cells of the (primary) coarse grid (white lines) serve as control volumes to build a coarse problem that defines the coarse-scale unknown at the nodes of the dual grid, or centers of the coarse grid (blue circles). Once the coarse solution is obtained, the interpolator can be used to obtain an approximate fine-scale solution.

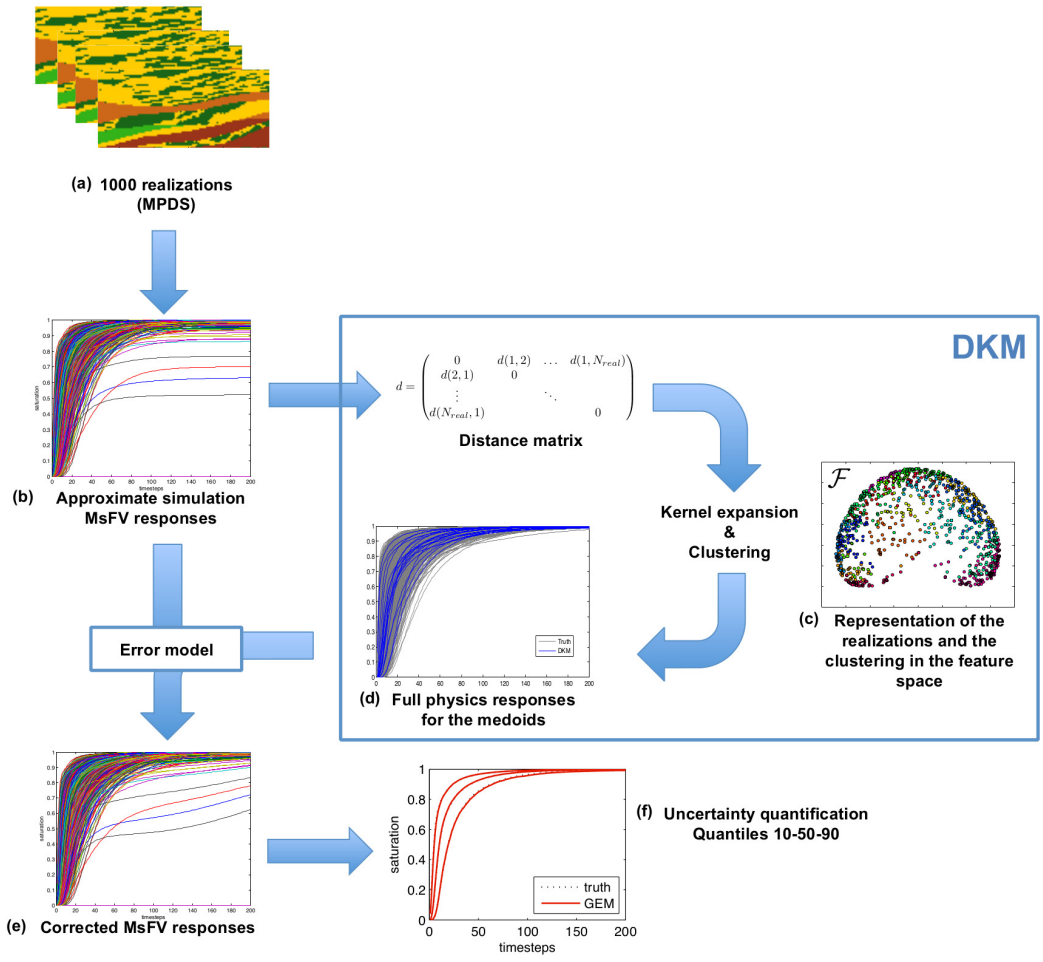
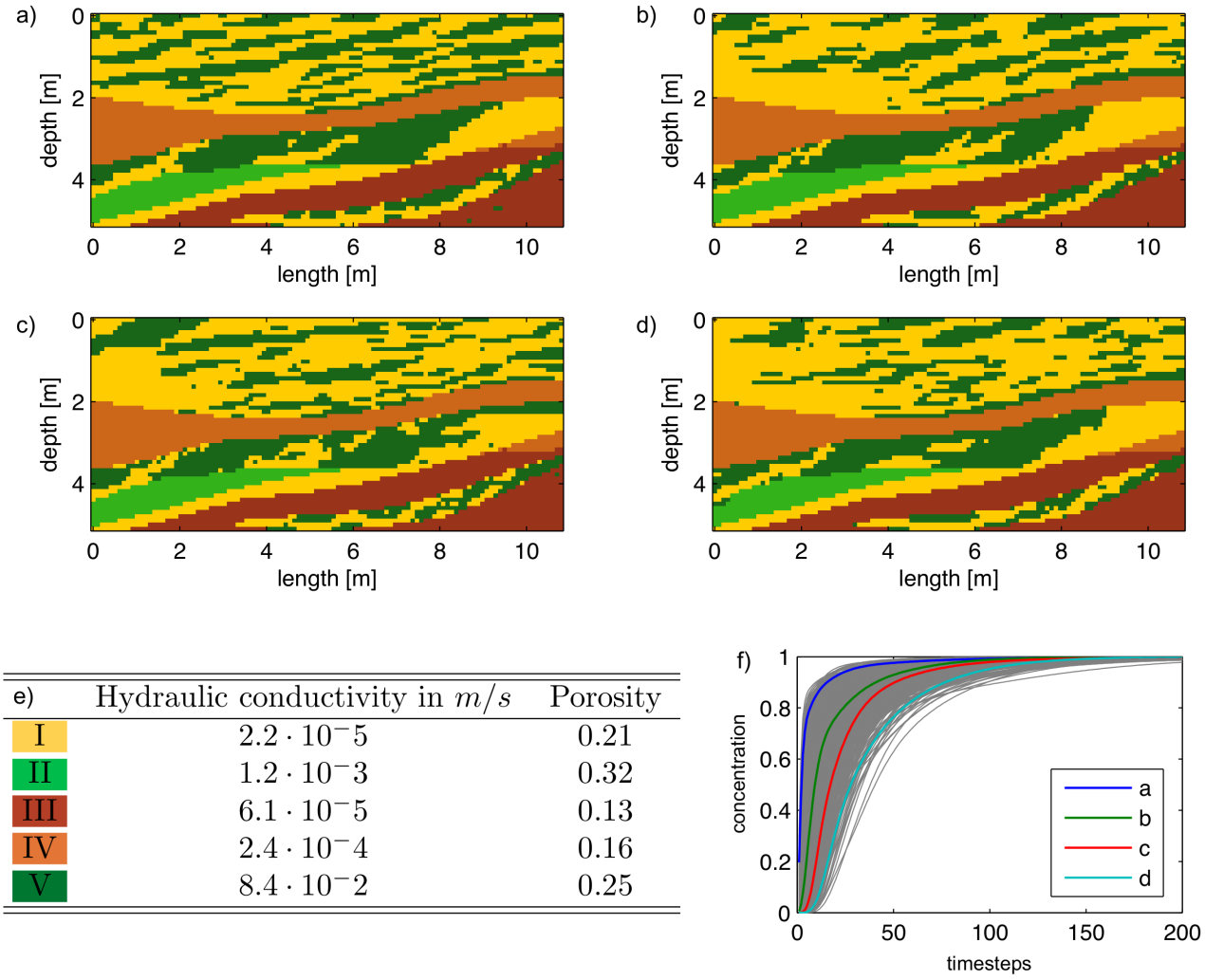


Figure 2: Starting from an set of geostatistical realizations $\{K_i, \phi_i\}$, MsFV simulations are run to compute a set of breakthrough curves, $\{C_i^{ms}(t)\}$. To select a subset of realizations, the euclidean distance between the curves, Eq. (7), is interpreted as a measure of dissimilarity. After the distance matrix $\mathbf{d}$ (Eq. 7) is constructed, a kernel method is used to compute a new distance in a feature space, $\mathbf{d}^{\mathcal{F}}$, Eq. (8). Based on $\mathbf{d}^{\mathcal{F}}$ the k-medoid algorithm, Eqs. (10) and (11), is used to cluster the realizations and finds a representative realization for each cluster (the medoid). After exact breakthrough curves are obtained for the medoids, the error model is constructed and generates the corrected curves, $\{C_i^*(t)\}$, which are used to compute the experimental quantiles.



e)	Hydraulic conductivity in m/s	Porosity
I	$2.2 \cdot 10^{-}5$	0.21
II	$1.2 \cdot 10^{-}3$	0.32
III	$6.1 \cdot 10^{-}5$	0.13
IV	$2.4 \cdot 10^{-}4$	0.16
V	$8.4 \cdot 10^{-}2$	0.25

Figure 3: (a-d) Examples of stochastic fields generated by DS (Mariethoz et al. 2010); (e) Table of the hydraulic conductivity and the porosity of the 5 lithofacies; (f) Breakthrough curves of the whole set of realization (grey) and of the four fields depicted in a, b, c and d (colors).

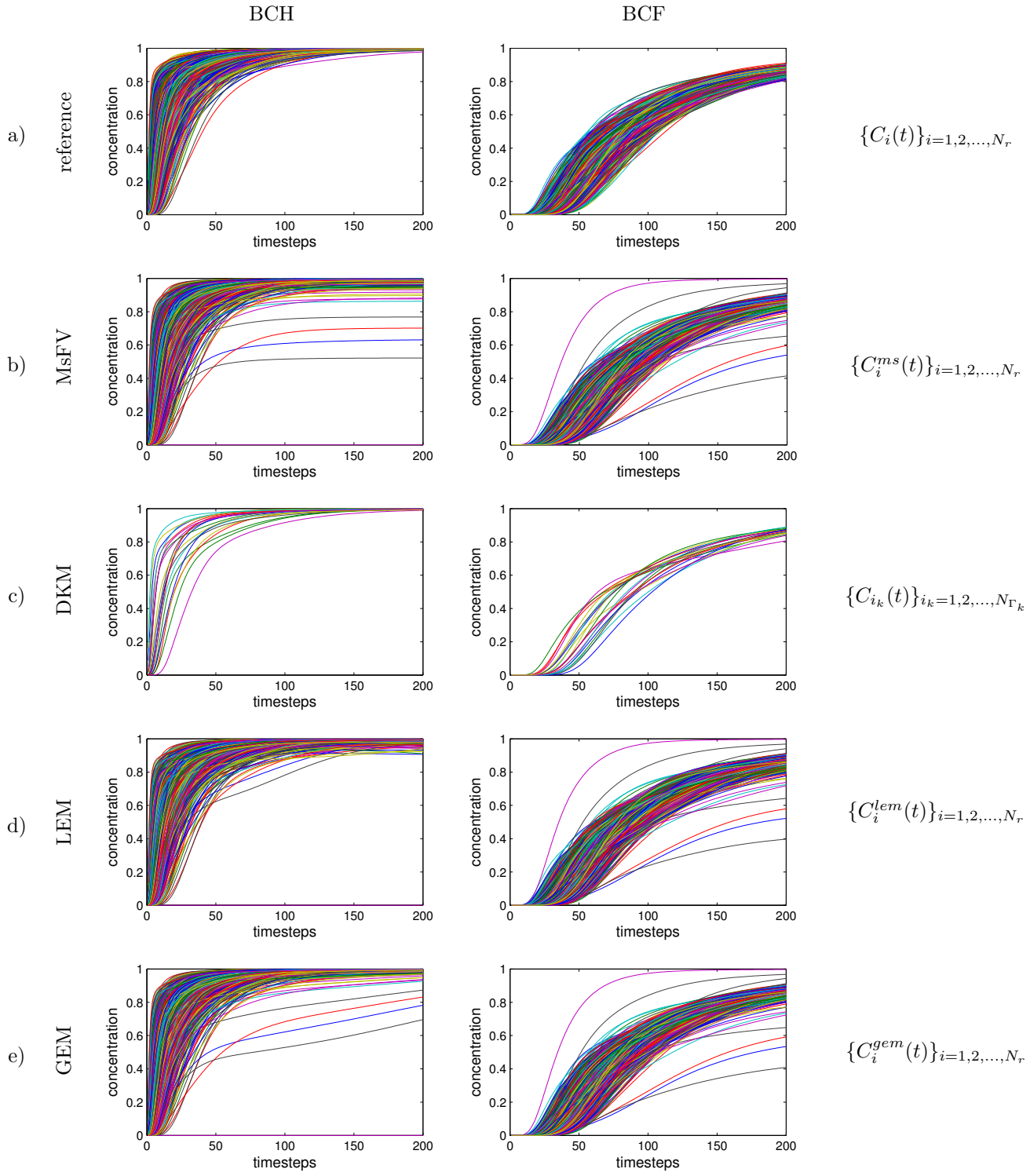


Figure 4: Ensemble of the breakthrough curves corresponding to each model, for the two types of boundary conditions: BCH (left), and BCF (right).

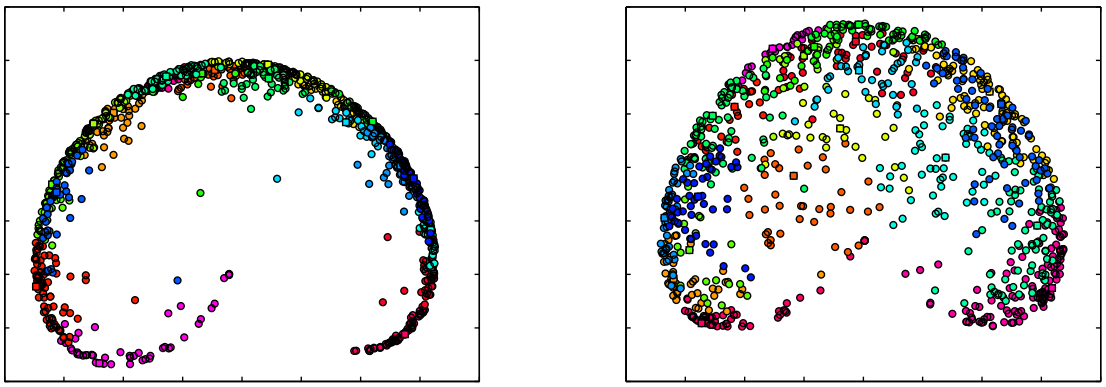


Figure 5: A two-dimensional representation of the feature space for BCH on the left and BCF on the right. Each dot is a realization, and each color represent a cluster. The realizations represented by a square are the medoids defined by the clustering algorithm, for which exact simulation are run. Note that this representation is obtained by the Multidimensional Scaling (Borg and Groenen 2005; Cox and Cox 2008; Scheidt and Caers 2009a), which is used here for visualization purposes only.

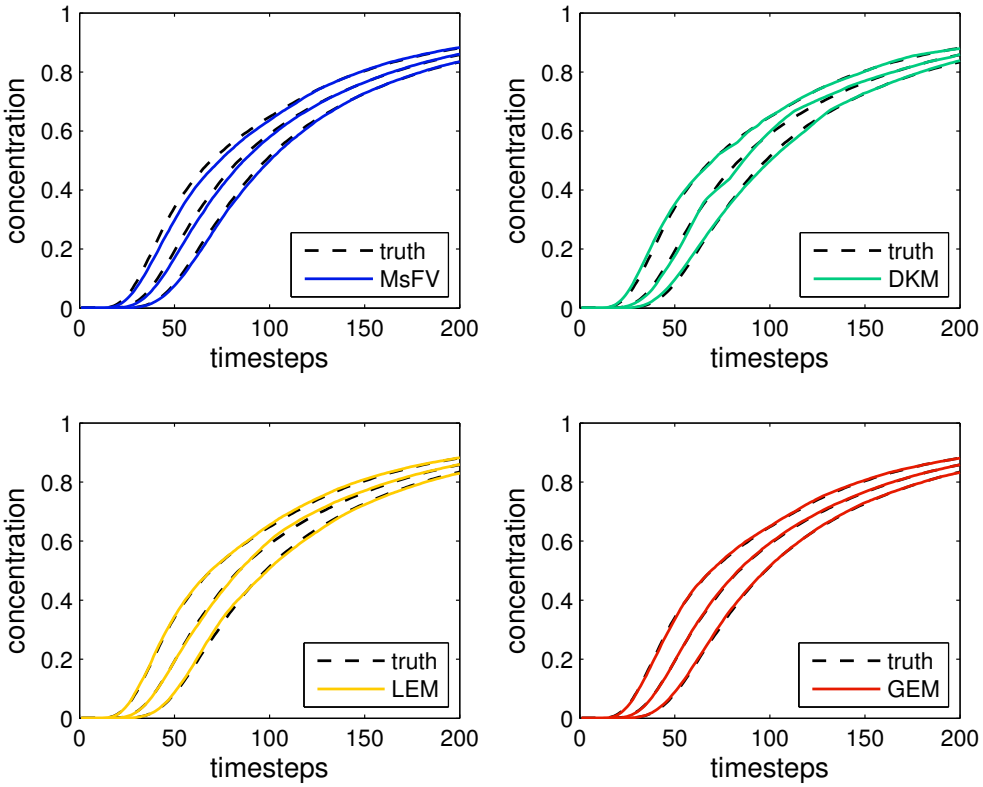


Figure 6: Quantiles curves estimated by MsFV (blue), DKM (green), LEM (yellow), and GEM (red) for the BCF.

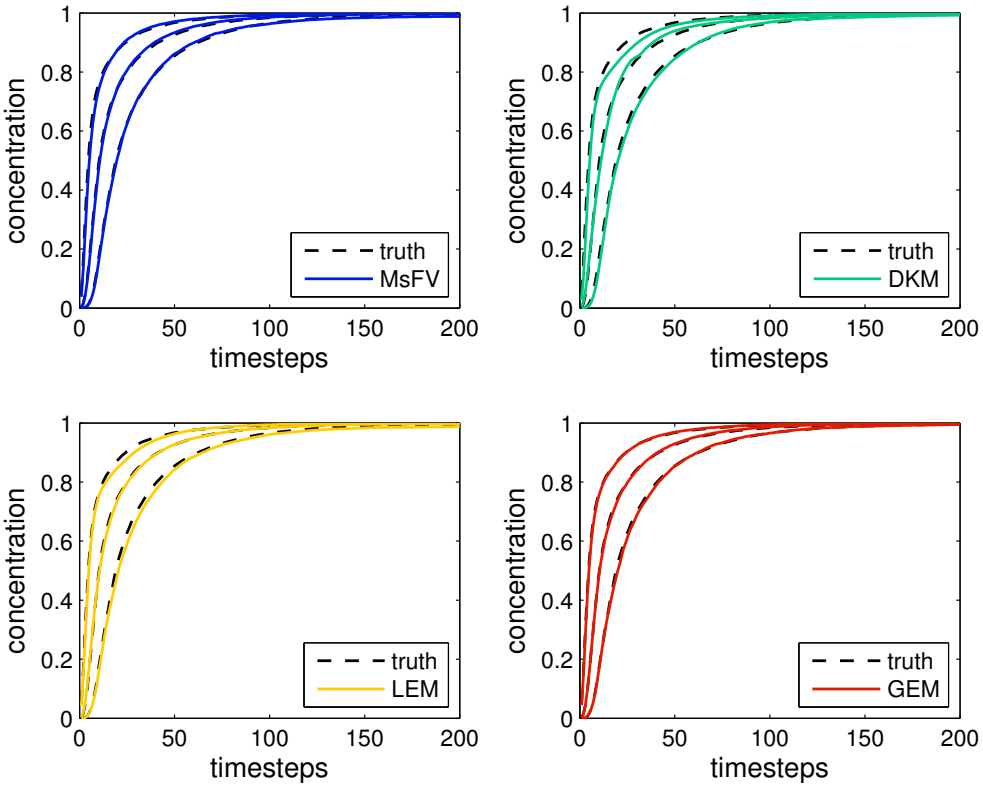


Figure 7: Quantiles curves estimated by MsFV (blue), DKM (green), LEM (yellow), and GEM (red) for BCH

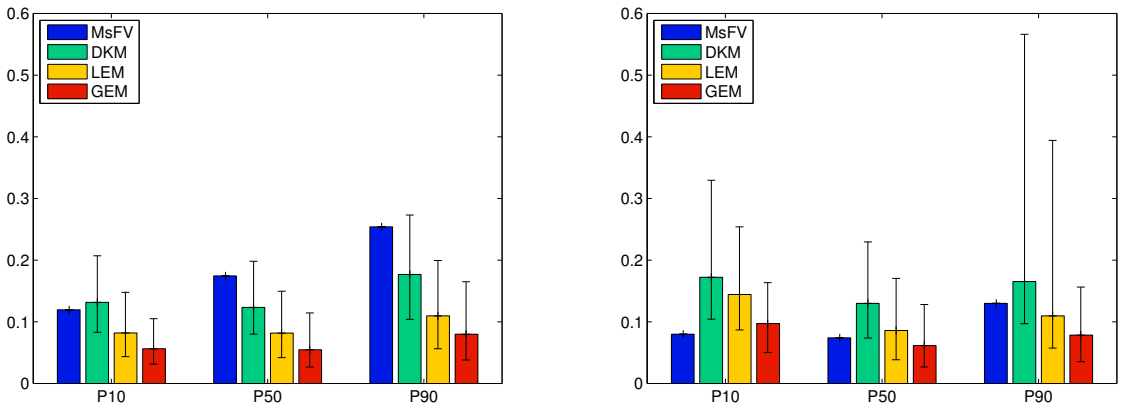


Figure 8: Errors on the quantile curves measured by the l_2-norm between the models and the reference curves. The bar plots represent the mean error of each method for each quantile curve. The error bars show the 80% confidence interval obtained for 500 results computed with different seeds. BCF is shown on the left and BCH on the right. Results for the l_∞-norm are shown in Fig. 9.

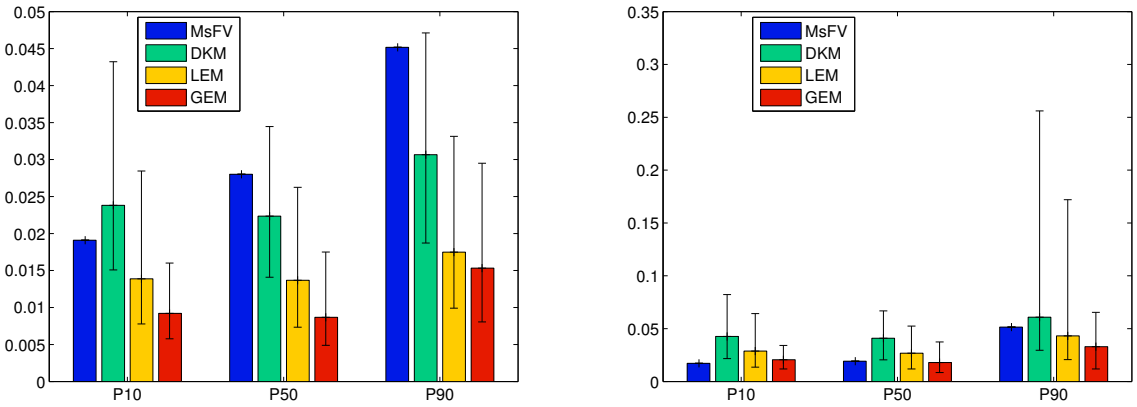


Figure 9: Errors on the quantile curves measured by the l_∞-norm between the models and the reference curves, for BCF (left) and BCH (right).

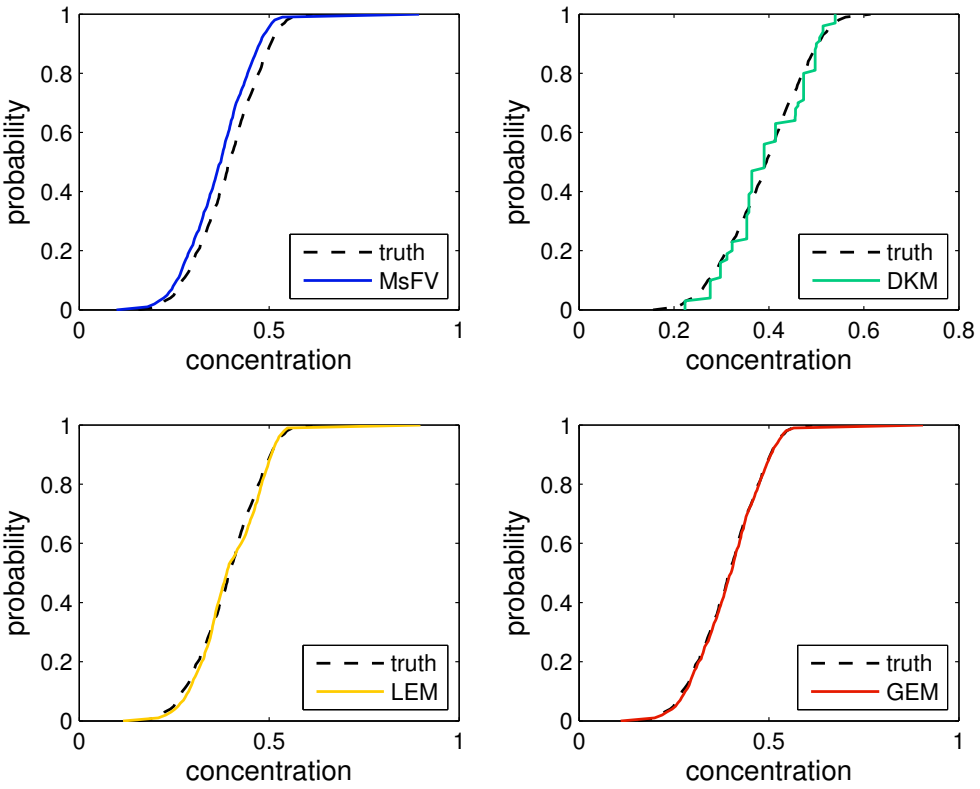


Figure 10: CDF of contaminant concentration at time step t = 70 for BCF.

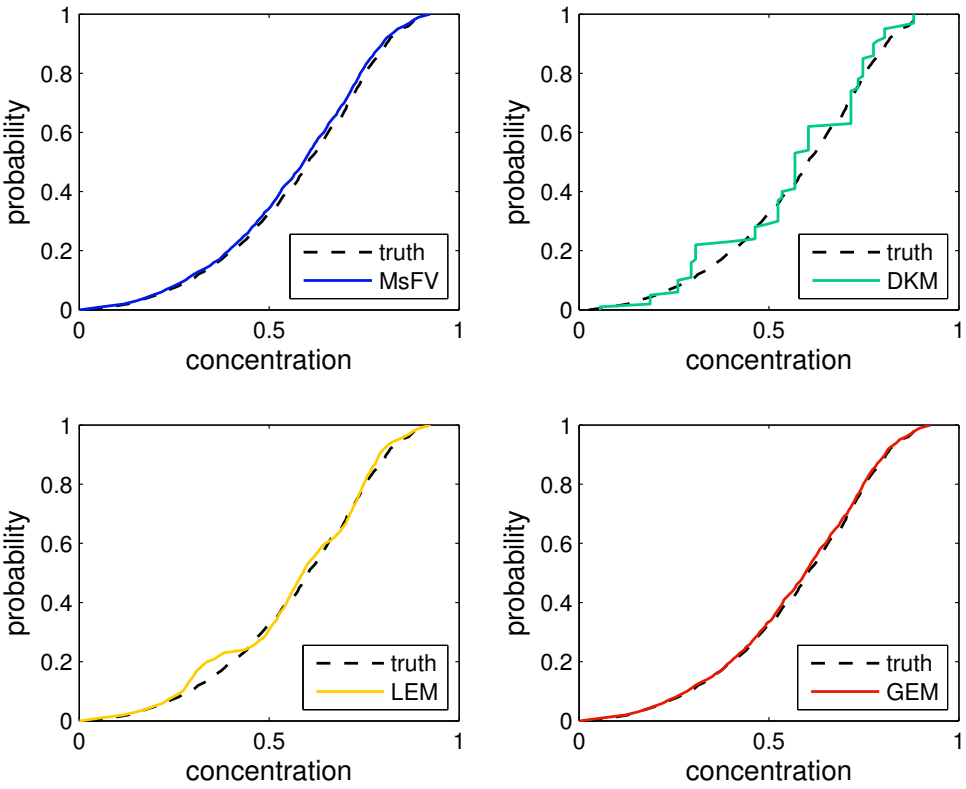


Figure 11: CDF of contaminant concentration at time step t = 14 for BCH.