Math Geosci (2009) 41: 397–419
DOI 10.1007/s11004-008-9186-0

Representing Spatial Uncertainty Using Distances and Kernels

Céline Scheidt · Jef Caers

Received: 2 July 2007 / Accepted: 17 June 2008 / Published online: 24 September 2008

Abstract Assessing uncertainty of a spatial phenomenon requires the analysis of a large number of parameters which must be processed by a transfer function. To capture the possibly of a wide range of uncertainty in the transfer function response, a large set of geostatistical model realizations needs to be processed. Stochastic spatial simulation can rapidly provide multiple, equally probable realizations. However, since the transfer function is often computationally demanding, only a small number of models can be evaluated in practice, and are usually selected through a ranking procedure. Traditional ranking techniques for selection of probabilistic ranges of response (P10, P50 and P90) are highly dependent on the static property used. In this paper, we propose to parameterize the spatial uncertainty represented by a large set of geostatistical realizations through a distance function measuring "dissimilarity" between any two geostatistical realizations. The distance function allows a mapping of the space of uncertainty. The distance can be tailored to the particular problem. The multi-dimensional space of uncertainty can be modeled using kernel techniques, such as kernel principal component analysis (KPCA) or kernel clustering. These tools allow for the selection of a subset of representative realizations containing similar properties to the larger set. Without losing accuracy, decisions and strategies can then be performed applying a transfer function on the subset without the need to exhaustively evaluate each realization. This method is applied to a synthetic oil reservoir, where spatial uncertainty of channel facies is modeled through multiple realizations generated using a multi-point geostatistical algorithm and several training images.

C. Scheidt (✉) · J. Caers
Department of Energy Resources Engineering, Stanford University, 367 Panama Street, Green Earth Sciences 353, Stanford, CA 94305-2220, USA
e-mail: scheidtc@stanford.edu

J. Caers
e-mail: jcaers@stanford.edu

Keywords Uncertainty quantification · Distance · Kernel methods · Ranking · Geostatistics

1 Introduction

Stochastic spatial simulation is now widely used to generate multiple, alternative realizations or samples of the same underlying spatial phenomenon, representing the uncertainty of the simulated variable(s). This set of realizations models the so-called "space of uncertainty" of the underlying phenomenon. In most applications, these realizations are not sufficient to assess uncertainty; further processing must be applied to address the practical questions at hand (Journel and Alabert 1990). For example, in reservoir engineering, several realizations of petrophysical and/or lithological models of the subsurface reservoir are generated and then submitted to flow simulation to assess reservoir flow performance, to assess the impact of drilling new wells or to optimize their placement. A similar situation arises in the management of groundwater, where subsurface model realizations are used to assess the impact of pumping on groundwater flow.

In general, a "transfer function" (e.g., flow simulator) is applied to post-process each realization and thereby obtain its "response" (e.g., reservoir performance), which may be single-valued or consist of a time-varying response. If many realizations are processed through the same transfer function, a probability distribution of the response can be constructed and serve as a model of uncertainty (Fig. 1). While this Monte Carlo simulation framework appears general and straightforward, several challenges make it difficult to apply. First, the set of realizations may be generated by varying several key parameters impacting spatial variation. Applying a single

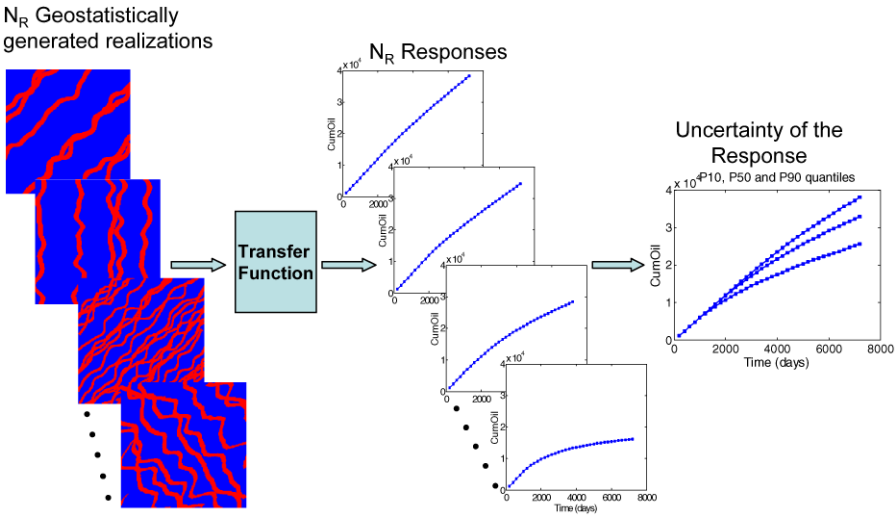


Fig. 1 Schematic diagram showing the calculation of probability quantiles though the evaluation of many realizations using a transfer function

geostatistical algorithm with a fixed parameter input often does not cover a wide enough space of uncertainty. However, by jointly varying several input parameters (variogram, histogram, training image, etc.) or even by varying the generating algorithm itself, one may need to create several hundred or thousand of realizations to capture the possible space of uncertainty adequately. Second, the transfer function may be expensive to evaluate in terms of CPU. Many transfer functions are either finite difference or finite element codes that may require some form of iterative optimization which may take several CPU hours if the grid underlying the realizations contains a large (10^5, 10^6) number of cells. It is therefore impractical, in most cases, to process several hundreds or thousand of realizations.

To circumvent these challenges, a widely used approach for modeling uncertainty is the experimental design technique (Box and Draper 1975). Experimental design aims at optimally selecting values of uncertain parameters in their range of variation, and then applying the transfer function on the resulting realizations to obtain responses (Damslet et al. 1992; Manceau et al. 2001). From the response values, a proxy model of the transfer function is built, which is a function of the uncertain parameters. The proxy model permits traditional Monte Carlo analysis of uncertainty. However, one major drawback of experimental design techniques is that they are often based on a simple linear regression and are thus not well suited for spatial variables or high-dimensional problems. In addition, experimental design techniques are not appropriate for applications with many discrete parameters.

An alternative approach to quantifying uncertainty is to examine a large set of realizations, and not individual parameters as done in experimental design. For problems of large dimensionality, evaluating the uncertainty of each parameter separately is often not useful, since many parameters are correlated, frequently in complex fashions. Moreover, ultimately, we are not interested in uncertainty of individual parameters, but in the realizations built from these parameters and responses predicted from those realizations. Ranking techniques are traditionally used to select realizations that represent the P10, P50 and P90 quantiles of the responses of interest (Ballin et al. 1992). The kth quantile is defined as the value x such that the probability of the response will be less than x is at most k% and the probability that the response will be less than or equal to x is at least k%. However, ranking techniques are highly dependent on the ranking property employed. Ranking is often based on a rather simple statistics extracted from the realization (e.g., original oil-in-place for reservoir models), which may not correctly capture the transfer function behavior. These statistical measurements often have a poor correlation with the response measured from the transfer function.

In this paper we propose a new approach which is well suited to treat problems using a very complex, time consuming transfer function, and non-Gaussian fields, which do not fit with traditional Least-Squares techniques. The key concept is the construction of a realization-based representation of uncertainty parameterized by distances. This method employs a single parameter (the distance) between any two realizations, which can be tailored to a particular application at hand. The aim is to select a set of representative realizations by analyzing the properties of the realizations as characterized by the distance, and finally quantify uncertainty within that set. Contrary to experimental design methodology which constructs a proxy model

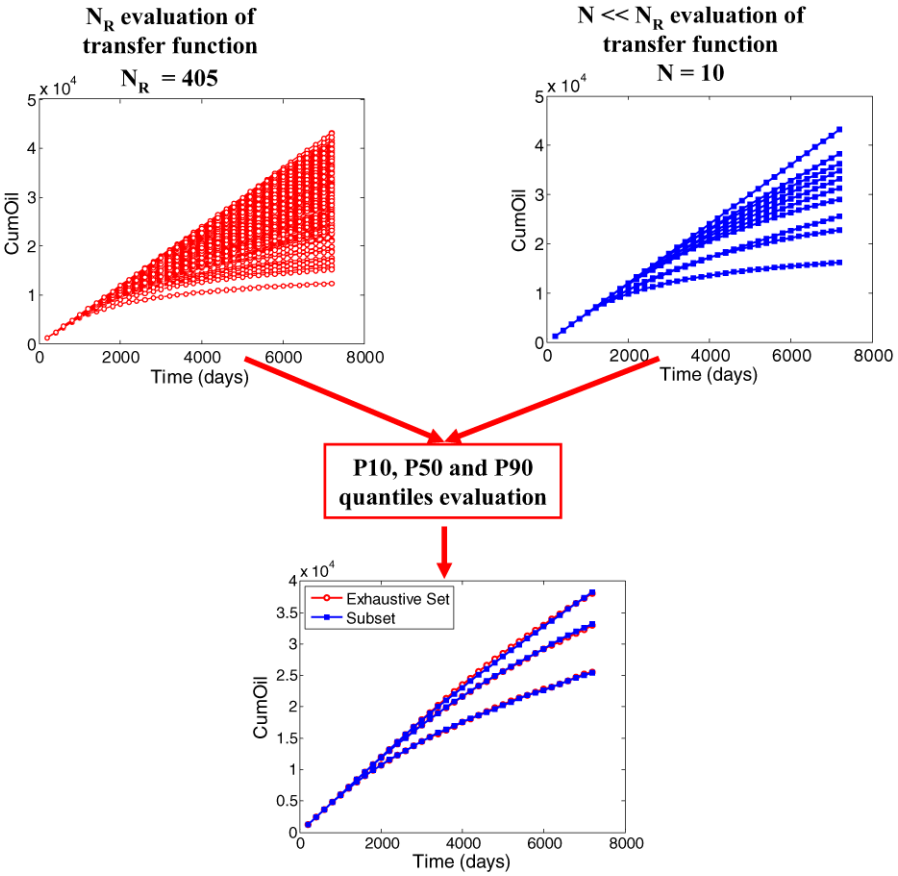


Fig. 2 Estimation of quantiles P10, P50 and P90 of a large set of responses by using only a few well selected realizations

of the response, the proposed method relies on the assumption that a few selected realizations have the same statistical characteristics in terms of response as the entire set (Fig. 2). Thus, no prediction outside of the existing set of responses is done with this method—the key is to select properly the subset of realizations. The following section describes the methodology used in this approach. An application of the methodology on a synthetic reservoir case of channel facies is presented. We end this paper by giving some conclusions and a discussion of future work.

2 Description of Methodology

The objective of the methodology is to efficiently select, among a potentially large set of realizations, a subset whose response (evaluated from a given transfer function) exhibits the same statistical properties (densities, quantiles, mean, variance, etc.) as the entire set of realizations. Since the transfer function can be very CPU demanding, the objective is to avoid evaluating realizations having similar responses, and to

concentrate on realizations which span a variety of response behavior. One way to do so is to attempt to quantify differences or similarities between realizations, and then group together realizations which are similar.

The principle of the methodology, illustrated for realizations of a binary categorical variable, is described in Fig. 3. Each step in Fig. 3 will be described in greater detail below. Starting with multiple (N_R) realizations generated using any algorithm, a dissimilarity distance matrix is constructed (Fig. 3(A) and 3(B)). This $N_R \times N_R$ matrix contains the "distance" between any two realizations. The matrix is then used to map all realizations into a Euclidean space $\boldsymbol{R}$ (Fig. 3(C)) using multidimensional scaling. Each point in this map represents a realization. Since in most cases the structure of the points in mapping space $\boldsymbol{R}$ is not linear, we use kernel methods to transform the Euclidean space $\boldsymbol{R}$ into a new space $\boldsymbol{F}$, called the feature space (Fig. 3(D)). The goal of the kernel transform is that points in this new space behave more linearly, so that standard linear tools for pattern detection can be used more successfully (Principal Component Analysis, cluster analysis, dimensionality reduction, etc.). These tools allow the selection of a few "typical" points representing realizations, among a potentially very large set. Application of the transfer function on a small subset of realizations allows uncertainty quantification (e.g., P10, P50, P90 quantiles) of the response variable.

2.1 Measurement of Dissimilarity Distance

2.1.1 Definition of Distance

The first step of the methodology is the definition of a dissimilarity distance between any two realizations (Fig. 3(A)). The concept of similarity between geostatistical model realizations was introduced by Arpat (2005), and Suzuki and Caers (2008). The distance is a way to determine how similar two realizations are in terms of spatial properties and transfer function response. The distance between two realizations can be determined by classical distances that measure difference in geometry such as the Hausdorff distance (Dubuisson and Jain 1994). For simulation of categorical variables, the Hausdorff distance focuses on foreground facies pixels (if binary model) or edge pixels of binary edges extracted from any type of realizations (Suzuki and Caers 2008). However, the Hausdorff distance does not take into account connectivity between wells which may be necessary to evaluate properly differences in flow. Connectivity-based distances (Park and Caers 2007), or streamline-based simulation are other examples of possible distances. Note that the concept of "relative" distance between two "objects" is different from the concept of difference in "absolute" statistical summaries. It is not necessary to determine statistical measures for each realization to measure dissimilarity. We need only to define a distance between any two realizations. The dissimilarity distance can be measured in any fashion; the only requirement for the distance between two realizations is that it should have a reasonable correlation with the difference in response of the same two realizations.

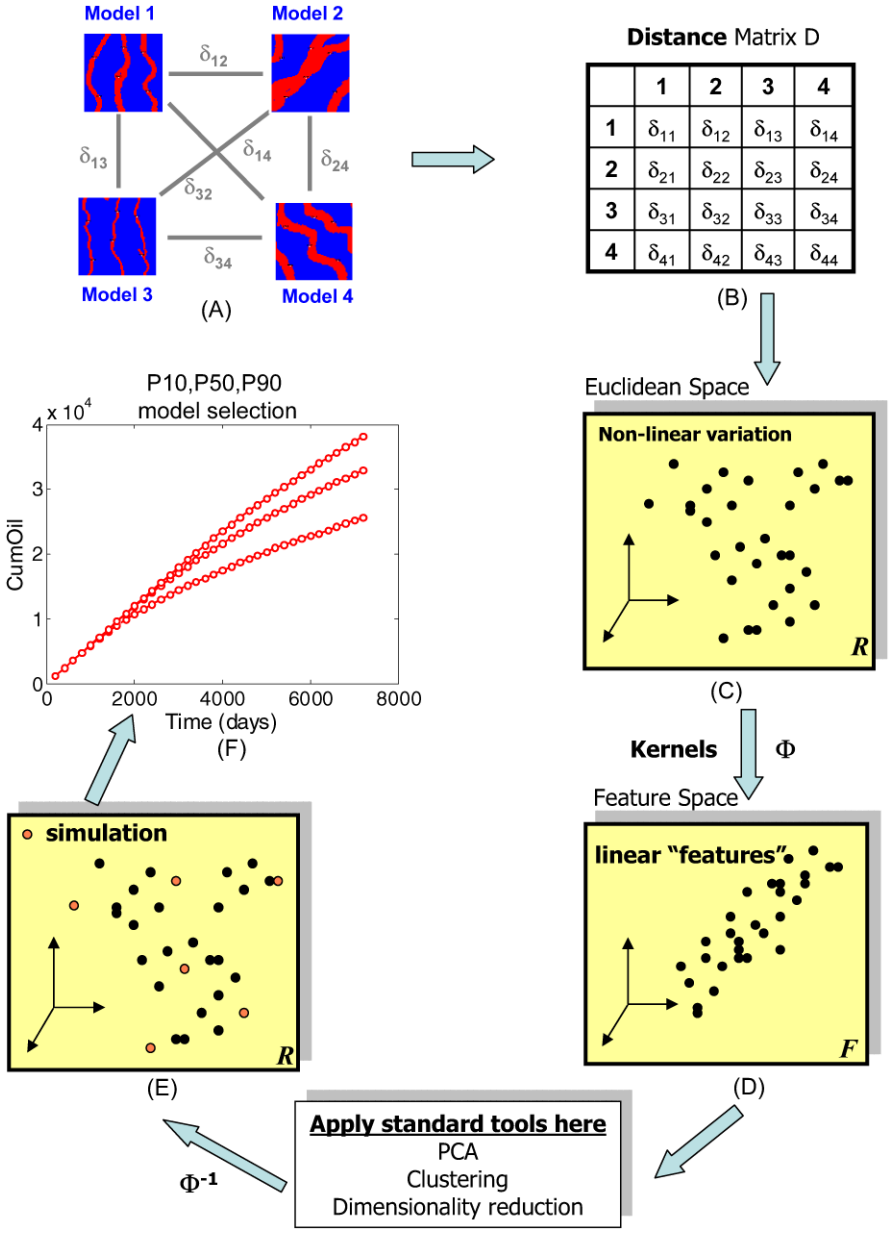


Fig. 3 Proposed workflow for uncertainty quantification—(**A**) distance between two models, (**B**) distance matrix ***D***, (**C**) models mapped in Euclidean space, (**D**) feature space, (**E**) pre-image construction, (**F**) P10, P50 and P90 quantile estimations

2.1.2 *Construction of a Dissimilarity Distance Matrix*

Given a set of N_R realizations $\boldsymbol{y}_i$ and a distance function δ between any two realizations, a $N_R \times N_R$ dissimilarity distance matrix $\boldsymbol{D}$ is constructed containing the distance measured between any two realizations δ_{ij}. A valid dissimilarity matrix must satisfy both of the following constraints: self-similarity ($\delta_{ii} = 0$) and symmetry ($\delta_{ij} = \delta_{ji}$). Once the distance matrix $\boldsymbol{D}$ is constructed, all the N_R realizations are mapped into an Euclidean space $\boldsymbol{R}$ using multidimensional scaling (MDS).

2.2 Multidimensional Scaling (MDS)

MDS is a technique used to translate the dissimilarity matrix into a configuration of points in nD Euclidean space (Borg and Groenen 1997; Cox and Cox 1994). The points in this spatial representation are arranged in such a way that their Euclidean distances correspond as much as possible (in least square sense) to the dissimilarities of the objects. A successful MDS procedure results in a good correlation between the Euclidean distance and the dissimilarity distance. The classical MDS algorithm rests on the fact that the coordinate matrix $\boldsymbol{X}$ of the points can be derived by eigenvalue decomposition from a matrix $\boldsymbol{A}$ obtained by converting the dissimilarity matrix $\boldsymbol{D}$ into a scalar product. Note that since the map obtained by MDS is derived solely by the dissimilarity distances in the matrix, the absolute location of the points is irrelevant. The map can be subject to translation, rotation, and reflection, without impacting the methodology. Only the distances in mapping space $\boldsymbol{R}$ are of interest.

Applying this concept to our methodology, the objects under consideration are realizations of a spatial phenomena, thus each realization is represented as a point. MDS allows to represent each realization $\mathbf{y}_i \in R^{N_C}, i = 1, \ldots, N_R$ (defined by potentially millions of grid-blocks) in a reduced coordinate system $\mathbf{x}_i \in R^p, i = 1, \ldots, N_R$. The dimension p of the Euclidean space is defined according to the eigenvalue decomposition of $\boldsymbol{A}$. In most cases, p can be small, since only a few eigenvalues in the decomposition are significant. A higher dimensional space could be used to increase the correlation, however, it would be improved only slightly while making the convergence of the pre-image more difficult (see below for the description of the pre-image).

An illustration of an application of classical MDS is presented in Fig. 4, illustrated for facies models and using the Hausdorff distance to construct the dissimilarity matrix. In this example, a 2D Euclidean space is sufficient to obtain a good quality mapping. The Euclidean distances between any two points are very close to the distances in the dissimilarity matrix (correlation of 0.79). Figure 5 shows the increase of the correlation as a function of the dimension of the Euclidean space. Note that the use of high dimensional Euclidean spaces may not increase the correlation significantly. For most applications, the structure of the points in the mapping space $\boldsymbol{R}$ is not linear and therefore standard tools for pattern detection are not appropriate. To avoid this problem, we employ kernel methods.

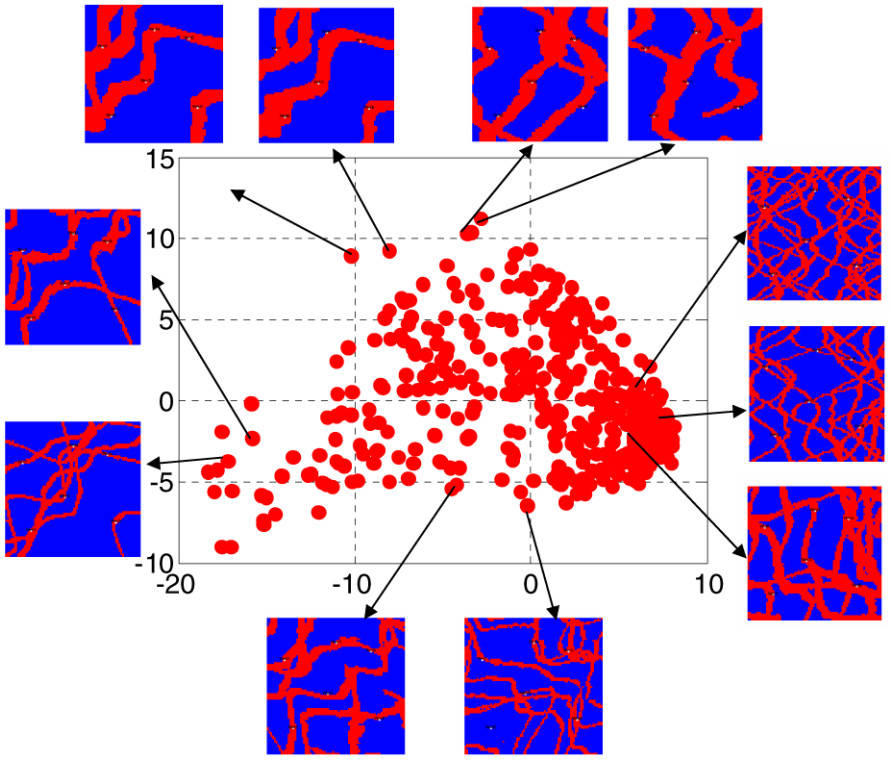
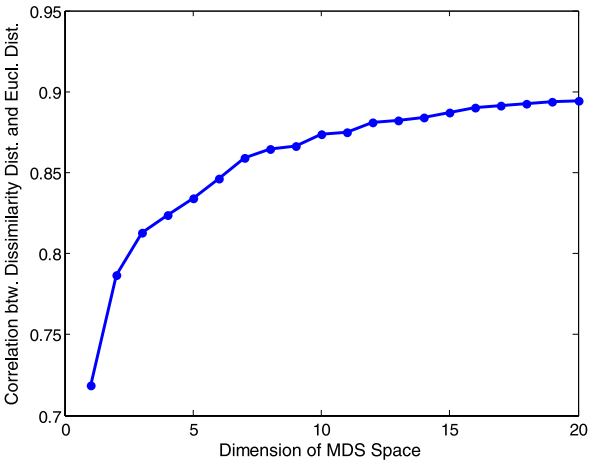


Fig. 4 Multidimensional Scaling (MDS): each point represents a reservoir model in a 2D space

Fig. 5 Correlation between dissimilarity distance and Euclidean distance as a function of the dimension of the Euclidean space



2.3 Kernel Methodology

2.3.1 Kernel-Principle

Kernel theory was recently developed in the field of neural computing and pattern recognition (Vapnick 1998). Kernel principal component analysis (KPCA) is often used as a tool to remove noise from computerized images (Shawe-Taylor and Cristianni 2004). In reservoir engineering, kernel theory has been used by Sarma (2006) in the context of inversion of flow data and production optimization. Kernel methods consist of mapping the given data points from their input space $\boldsymbol{R}$ to some high-dimensional feature space $\boldsymbol{F}$ using a multidimensional function $\Phi : \Phi : R \rightarrow F$. The feature space $\boldsymbol{F}$ is assumed to have a better linear variation than $\boldsymbol{R}$. In other words, points in $\boldsymbol{F}$ are linearly separable. Thus, tools requiring a linear relationship between data can be applied into $\boldsymbol{F}$ instead of $\boldsymbol{R}$. In our application, points generated by MDS in space $\boldsymbol{R}$ are transformed by kernel methods into space $\boldsymbol{F}$.

Kernel methods can be used to develop nonlinear generalizations of any algorithm that can be cast in term of scalar products, such as PCA or k-means clustering (Schöelkopf et al. 1998; Schöelkopf and Smola 2002). One principle advantage of using kernels in these applications is that there is no need to map explicitly the points from space $\boldsymbol{R}$ to $\boldsymbol{F}$; all necessary computations in space $\boldsymbol{F}$ can be carried out using the scalar product of the nonlinear function Φ. This function is called a kernel function k, and is given by

$$k(\mathbf{x}, \mathbf{y}) = \big\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \big\rangle. \tag{1}$$

The most common kernel function, the scalar product in the feature space $\boldsymbol{F}$, is the Gaussian kernel (radial basis function), which is given by

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} \right) \quad \text{with } \sigma > 0. \tag{2}$$

In our application, we consider a Gaussian kernel for all the cases (2). Tests on other kernel functions such as polynomial or sigmoidal were made, but the Gaussian kernel was found to be more robust. The kernel width parameter σ controls the flexibility of the kernel. From Keerthi and Lin (2003), we know that for small values of σ, the kernel matrix becomes close to identity matrix ($\boldsymbol{K} = \boldsymbol{I}$), and thus the application of KPCA will lead to severe over-fitting. On the other hand, large values of σ gradually reduce the kernel to a constant function ($\boldsymbol{K} = \mathbf{1}$). In this case, the application of KPCA will lead to severe under-fitting. As recommended by Shi and Malik (2000), we choose σ as 10% to 20% of the range of the distance between points. Robustness in the results of kernel clustering were found using σ within this range.

2.3.2 Pre-Image Construction

While the mapping Φ from input space $\boldsymbol{R}$ to feature space $\boldsymbol{F}$ is of primary importance in kernel methods, the reverse mapping from feature space $\boldsymbol{F}$ back to input space $\boldsymbol{R}$ may be desired (Fig. 3(E)). This reverse mapping process is called the pre-image

problem. For example, one may want to map back to $\boldsymbol{R}$ the points (in space $\boldsymbol{F}$) projected by KPCA into a lower dimensional space. The difficulty with this procedure is that the mapping function Φ into $\boldsymbol{F}$ is unknown, nonlinear and non-unique, thus only approximate solutions are possible. In this work, approximate pre-images are found using the fixed-point iteration approach proposed by Schölkopf. This approach is essentially a gradient-based optimization technique. Note that the pre-image points do not necessarily correspond to a point in the original space $\boldsymbol{R}$. In this case, we take the closest existing point. For details about the pre-images algorithm, see Schöelkopf and Smola (2002).

2.3.3 Kernel Principal Component Analysis (KPCA)

To understand KPCA better, we first recall quickly the theory of principal component analysis (PCA). PCA consists of projecting linearly the data onto a lower-dimensional space. PCA provides a set of orthogonal axes, called principal components, obtained by solving the eigenvalue problem of the sample correlation matrix. A small number of principal components is often sufficient to describe the major trend in the data. KPCA works in a similar manner. First, kernels map the given data points from space $\boldsymbol{R}$ to space $\boldsymbol{F}$ using a multidimensional function $\Phi : \Phi : \boldsymbol{R} \rightarrow \boldsymbol{F}$ and then PCA is applied in $\boldsymbol{F}$. Using the kernel function (1), it can be shown (Schöelkopf and Smola 2002) that KPCA requires only an eigenvalue decomposition of the $N_R \times N_R$ kernel matrix $\boldsymbol{K}$ defined by

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j), \quad \mathbf{x}_i \in R,\ i = 1, \ldots, N_R.$$

KPCA is a powerful technique for extracting structure from potentially high-dimensional data sets with complex variability. KPCA can be seen as a way of removing noise from the points in $\boldsymbol{R}$.

2.3.4 Kernel K-Means Clustering (KKM)

Clustering algorithms are also applicable in feature space $\boldsymbol{F}$ and are suited to our problem. Cluster analysis aims to discover the internal organization of a dataset by finding structure within the data in the form of clusters. Hence, the data is broken down into a number of groups composed of similar objects. This methodology is widely used both in multivariate statistical analysis and in machine learning. Defining clusters consists in identifying an 'a priori' fixed number of centers and assign points to clusters with the closest center. The number of the cluster is defined by the user, depending on the number of evaluations of the transfer function which can be performed in the time allotted to the task. In this work, we apply the classical k-means algorithm in the feature space $\boldsymbol{F}$ to determine a subset of points defined by the cluster centroids. The k-means procedure requires a method for measuring the distance between two points in the high-dimensional feature space $\boldsymbol{F}$. The distance can be computed using the scalar product information through the equality

$$\begin{aligned} \left\| \Phi(\mathbf{x}) - \Phi(\mathbf{z}) \right\|^2 &= \left\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \right\rangle + \left\langle \Phi(\mathbf{z}), \Phi(\mathbf{z}) \right\rangle - 2\left\langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \right\rangle \\ &= k(\mathbf{x}, \mathbf{x}) + k(\mathbf{z}, \mathbf{z}) - 2k(\mathbf{x}, \mathbf{z}). \end{aligned}$$

Note that this equality is only true for Euclidean distance, hence the necessity of the MDS procedure prior to performing KKM. For an overview of clustering techniques, see Buhmann (1995), and Shawe-Taylor and Cristianni (2004) for specific information about kernel clustering techniques.

2.4 Uncertainty Quantification

In our application, we apply KPCA or KKM to hundreds of realizations of a spatial phenomenon mapped in a Euclidean space. These methods identify a small subset of realizations which represent "typical" realizations of the full set. In Fig. 6(A), we use the same example as in Fig. 4 using the Hausdorff distance—the subset of realizations selected by KPCA is represented by squares. Application of the transfer function (e.g., flow simulation) is then done on this subset of realizations. Uncertainty can be subsequently analyzed by calculating, for example, the quantiles P10, P50 and P90 on these few models (Fig. 6(B)). Note that the subset of realizations selected by KPCA or KKM may not be equiprobable. Thus, a weighting scheme should be defined for a proper estimation of the quantiles. In the case of KKM, it is obvious to represent each simulation as many times as the number of models in the corresponding cluster. In the case of KPCA, we propose to perform k-means in the subspace generated by KPCA to define the weighting scheme.

2.5 Illustration of the Concept in a Simple Example

Before providing a realistic application of this methodology, we give a simple illustrative example that describes intuitively the inner working of the proposed approach. In this example shown in Fig. 7, we have generated 30 2D-realizations $\mathbf{x}^i = (x_1^i, x_2^i)$, $i = 1, \ldots, 30$. Their representation in the 2D parameter space is given by 3 clusters as shown by the points, colored by cluster. Such clusters could represent three "geological populations". The transfer function for this example is defined by $f(x_1, x_2) = 0.5x_1 + x_2^3$. The contours of the values of the transfer function are shown

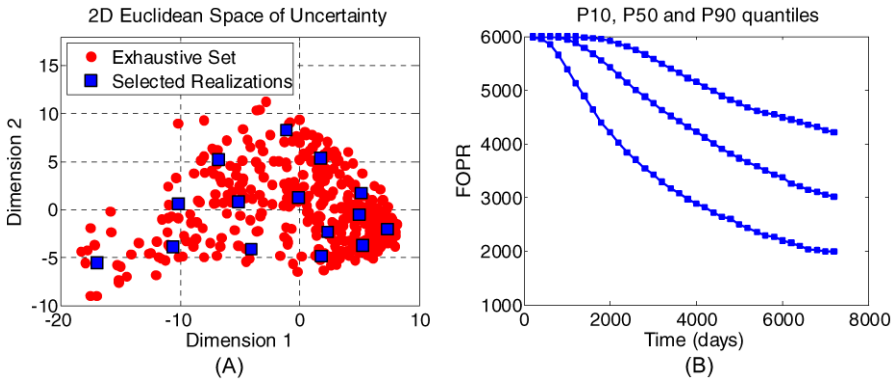


Fig. 6 (**A**) Mapping space R: points selected by KPCA represented by *squares*, (**B**) Resulting quantiles estimation (P10 P50 and P90)

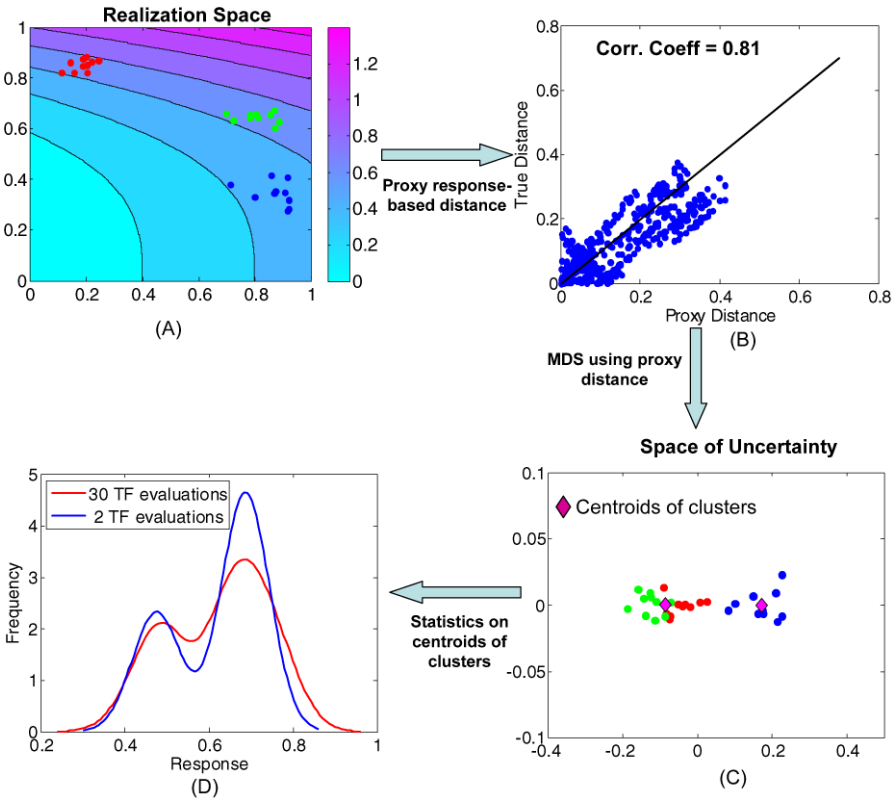


Fig. 7 Simple example demonstrating the benefits of using a good distance to group realizations having similar transfer function values

in Fig. 7(A). Note that in the general case, for CPU reasons, the transfer function cannot be evaluated exhaustively. To assess response uncertainty on these realizations, one could select the centroids of each cluster in Fig. 7(A) and evaluate the transfer function. This would result in 2 evaluations with similar responses, which should be avoided in the case of a CPU demanding transfer function.

However, assume we can define a distance between realizations i and j, which is defined as a weighted difference between the coordinates of the points (3).

$$d_{ij} = 0.5\left(x_1^i - x_1^j\right) + \left(x_2^i - x_2^i\right). \tag{3}$$

The distance is a 1st order approximation of the actual difference in transfer function. The correlation coefficient between d_{ij} and the actual distance is 0.81 (Fig. 7(B)). If we compute the distance between each pair of realizations, we can map all the realizations using MDS into a new 2D space as shown in Fig. 7(C). We can see that 2 groups are identified by traditional k-means in the MDS space, and only 2 transfer function evaluations are necessary. The probability density derived from these 2 transfer function evaluations reproduces accurately the density of the response of the 30 realizations (Fig. 7(D)). The probability densities were generated using a kernel smoothing

algorithm (Bowman and Azzalini 1997). This example, albeit simple, emphasizes an important point that applies generally. Spatial model realizations may exist in a high-dimensional space (e.g., the number of grid cells). Examining the responses in such a high-dimensional space may be prohibitive. However, certain realizations may appear different from a spatial/geological point of view but may have similar responses. Instead, if a response specific distance exists, the realizations can be mapped into a low-dimensional response space (Fig. 7(C)) allowing efficient and effective selection of representative realizations for quantifying response uncertainty. Note that, in the example above, the realizations in the response space behave linearly and clustering techniques can be applied directly in this space. However, in real applications, the Euclidean space resulting from the MDS procedure is often very complex because of the high non-linearity of the transfer function. Thus, kernel techniques are employed to make the clustering procedure more efficient. Results of the application of KPCA and KKM for an oil reservoir example are presented in next section.

3 Application to Subsurface Flow Uncertainty Assessment

In reservoir engineering, several realizations of petrophysical and/or lithological models of the subsurface reservoir are generated using a geostatistical algorithm. The transfer function in this application is a numerical flow simulator, which can be very CPU demanding for models with a large number of grid cells. The realizations are submitted to flow simulation in order to assess the uncertainty in reservoir flow performance. We use the method proposed in this paper to perform only a small number of flow simulations whose response has the same characteristics as the entire set of realizations. The probability distribution of the flow response of interest, in this case the field production oil rate, is then determined. A synthetic case study is presented to demonstrate the potential of the proposed methodology. We consider a channel system, composed of mud and sand. The background mud is understood as a sealing rock which does not have flow capacity and storage capacity. The inner channel heterogeneity in petrophysical properties is considered as negligible, thus channels are modeled with uniform (known) porosity/permeability. The spatial distribution of channel sands is considered the main driving parameter for flow. The reservoir model is a 80×80 2D grid containing 3 producers and 3 injectors, all penetrating channel sand. To construct a prior uncertainty space that accommodates a large set of model realizations, 5 realizations derived from 81 training images (giving a total of 405 facies realizations) are geostatistically simulated. The realizations are conditioned to facies observations at the wells, depicted in Fig. 8, all penetrating channel sand. A multipoint geostatistical technique called SIMPAT (Arpat and Caers 2007) is used. Note that in this case, we vary a key input component of the algorithm, namely the training image. Since the training image determines the nature of the pattern being simulated, the 405 realizations exhibit significantly varying patterns. In order to analyze the efficiency and quality of the method, standard flow simulations were run for each realization. Note that for real field cases, this is in general not possible. For a more detailed description of the case, see Suzuki and Caers (2008).

We apply the methodology proposed in this paper using flow-based distances measured using streamline simulation. The Hausdorff distance showed a poor correlation

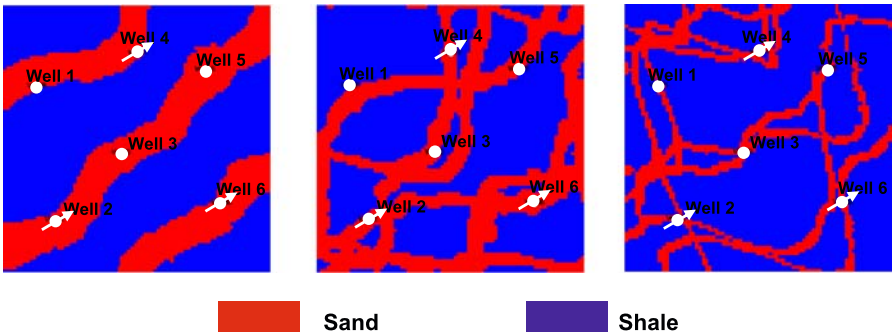


Fig. 8 Example of 3 reservoir model realizations with well locations

with the flow simulations, thus was not well suited to the application. We have found that flow-based distances are well adapted when the response of interest is production data, which is the case in this study. Streamline simulation has been shown to be orders of magnitude faster than standard flow simulation, and is thus well suited for problems where rapid evaluation of many models is needed (Batycky et al. 1997). Our results are compared with traditional methods, such as ranking with static properties and tracer simulations.

3.1 Application of the Proposed Methodology

3.1.1 Construction of the Distance Matrix

The distance is calculated using the tracer simulation option of a commercial streamline flow simulator. Tracer simulation approximates the flow as linear, meaning that the injection and production fluids are assumed to be identical. Thus, in this case, only a single pressure solve is necessary to perform fluid flow simulation, giving extremely rapid results. The distance between any two reservoir models is given as the absolute difference in field oil rate at two given times (10 000 and 20 000 days):

$$\delta_{ij} = \sum_{t\in\{10000,20000\}} \left|\mathrm{FOPR}_i^{\mathrm{streamline}}(t) - \mathrm{FOPR}_j^{\mathrm{streamline}}(t)\right|.$$

The field oil rate for each model differs due to the difference in water breakthrough for each producing well. Late simulation times are employed to ensure that the water breakthrough has occurred for all realizations, enabling the greatest distinction in the water breakthrough between each realization. As discussed before, the distance needs to be reasonably well correlated with the flow response of interest. To compare the distance with the results from standard flow simulation, we calculate the average of absolute difference in oil rate as

$$\Delta_{ij}^{\mathrm{FOPR}} = \frac{1}{N_t}\sum_{t=1}^{N_t} \left|\mathrm{FOPR}_i^{\mathrm{Eclipse}}(t) - \mathrm{FOPR}_j^{\mathrm{Eclipse}}(t)\right|, \tag{4}$$

where t represents the time and N_t the number of timesteps. In this case, the correlation coefficient between the distance and the difference in oil is $\rho(\delta, \Delta^{\mathrm{FOPR}}) = 0.77$ which we have found is generally sufficient for accurate results. A smaller correlation coefficient will not necessarily result in inaccurate uncertainty quantification; most of the time increasing the number of clusters to retain more simulations is sufficient.

3.1.2 Multi-Dimensional Scaling

Using the dissimilarity distance previously defined, we apply multidimensional scaling to map all the realizations in a 3D Euclidean space $\boldsymbol{R}$. A 3D mapping space $\boldsymbol{R}$ is deemed appropriate to ensure that the Euclidean distance between any two points in $\boldsymbol{R}$ reproduces the dissimilarity in the matrix $\boldsymbol{D}$. Indeed, the correlation coefficient between the dissimilarity matrix $\boldsymbol{D}$ and the pair-wise Euclidean distance is high at 0.9. Subsequent application of KPCA and KKM only considers these Euclidean distances between the models since Euclidean distance is a very good representation of the dissimilarities of the reservoir models. Note that distinct geological properties give rise to a wide scatter of flow responses. Note as well that certain realizations may appear different from a geological point of view, but they may exhibit similar flow behavior. In this case, these realizations would then be found in the same cluster, regardless which training image they were derived from.

3.1.3 Application of KPCA

At this step of the methodology, we define a kernel function which transforms the mapping space $\boldsymbol{R}$ into a space $\boldsymbol{F}$ with improved linear variation. The Gaussian radial basis kernel is used (2), whose parameter was chosen as 10% of the range of the dissimilarity matrix: $\sigma = 800$. We first perform KPCA to reduce the dimensionality of the problem by projecting the points in a 4D subspace of the feature space. In that subspace, we perform cluster analysis using k-means, to determine 15 clusters. The number of cluster was defined as the maximum number of flow simulations we can afford for a given CPU. We compute the pre-images of each centroid using the Schölkopf fixed-point algorithm.

Full flow simulations are performed for 15 realizations corresponding to the pre-images. Recall that the pre-image mapping may not result in points corresponding to a realization. In this instance, we select the nearest point (realization) for flow simulation. Uncertainty quantification is then performed by calculating the quantiles P10, P50 and P90 on these 15 models as a function of time, each model being represented as many times as the number of models in the corresponding cluster. Thus, only 15 flow simulations were performed out of a total of 405 reservoir models. Figure 9(A) represents the evolution of field oil rate as a function of time for the 15 selected realizations; Fig. 9(B) represents the quantiles resulting from the 15 and 405 simulations, respectively. We can observe that the estimation of quantiles P10, P50 and P90 is accurate. In Figs. 9(C) and 9(D), we present the probability density of the field oil rate for the 405 realizations (dotted line), and for the 15 realizations for 2 different times (2000 and 6000 days). Table 1 shows the mean, variance, kurtosis, and skewness coefficients of the densities. We can see that the estimated densities are

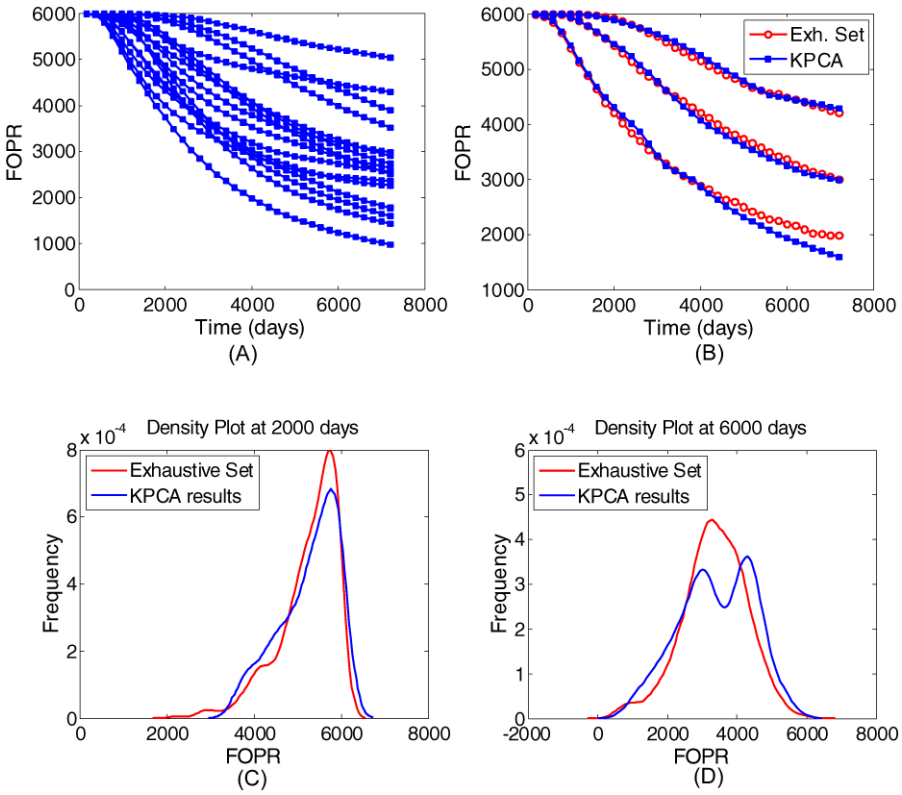


Fig. 9 KPCA Results: (**A**) Oil Rate as a function of time for all the 15 realizations, (**B**) P10, P50 and P90 values, (**C**) Density of oil rate for all 405 realizations (*dashed line*) and 15 selected realizations (*solid line*) at 2000 days and (**D**) at 6000 days

Table 1 Statistical properties of the densities resulting from KPCA and KKM

	All Data 2000 days	KPCA 2000 days	KKM 2000 days	All Data 6000 days	KPCA 6000 days	KKM 6000 days
Mean	4.11E+03	4.84E+03	4.35E+03	3.25E+03	3.20E+03	2.63E+03
Variance	2.04E+06	1.01E+06	1.25E+06	4.29E+06	1.53E+06	2.07E+06
Kurtosis	1.7998	1.7998	1.7998	1.7998	1.7998	1.7998
Skewness	8.71E−16	5.51E−16	−2.99E−15	−1.74E−16	−1.13E−15	1.24E−15

close to the reference, which means that the 15 realizations have similar characteristics as the 405. We now replace this two-step methodology (PCA and clustering) by performing a single cluster analysis in the feature space $\boldsymbol{F}$.

3.1.4 Application of KKM

In this section, we apply the second methodology, kernel k-means (KKM). The two first steps, the definition of the dissimilarity matrix and mapping the points with MDS, are identical for the 2 methods. Thus, they are not illustrated here. A Gaussian kernel (2) with $\sigma = 850$ is used to define the feature space $\boldsymbol{F}$ in which clusters are identified. Again, we assume that only 15 flow simulations are affordable in this case. Uncertainty quantification is subsequently performed by flow simulation for the 15 realizations selected by KKM and by computing the resulting quantiles P10, P50 and P90 (Figs. 10(A) and 10(B)). The weight of each realization equals the number of points in the corresponding clusters. Figures 10(C) and 10(D) represent the density computed from the 15 simulations, as well as the density for the full set of realizations for 2 different times. Table 1 gives their statistical properties. We can see that the subset of 15 selected realizations has similar probability densities compared to the entire set of 405 realizations.

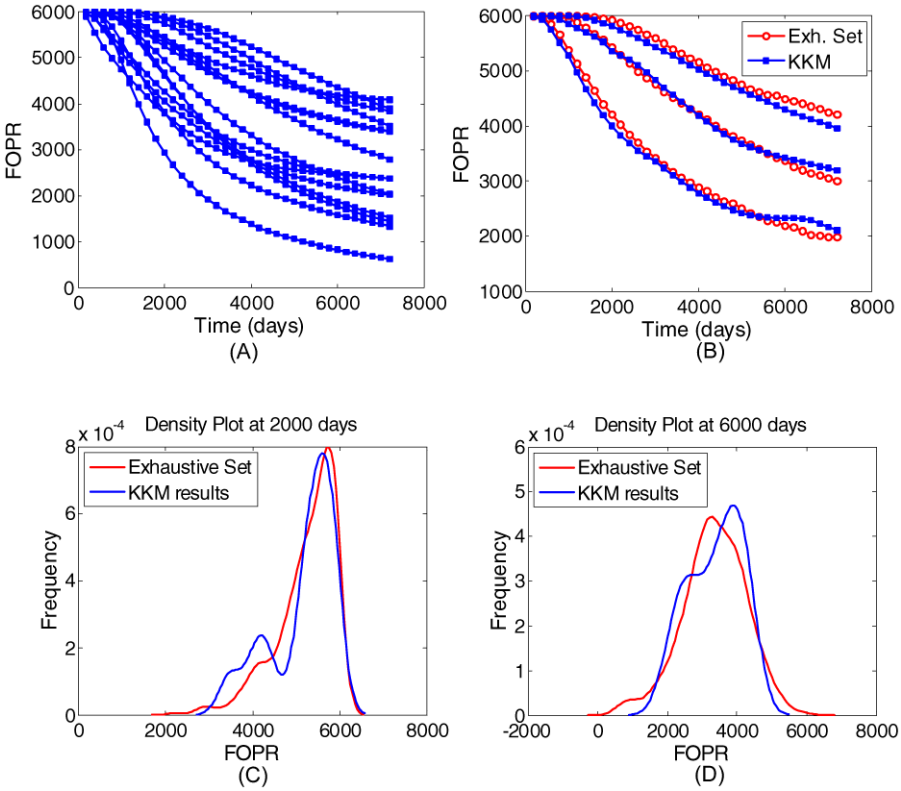


Fig. 10 KKM Results: (**A**) Oil Rate as a function of the time for all the 15 realizations, (**B**) P10, P50, P90 values, (**C**) Density of oil rate for all 405 realizations (*dashed line*) and 15 selected realizations (*solid line*) at 2000 days and (**D**) at 6000 days

3.2 Comparison with Classical Ranking Techniques

In this section, we compare the results obtained by the proposed methodology with classical techniques which consists of ranking the realizations according to a specific measure, and then determining realizations for flow simulation.

3.2.1 Ranking Technique Review

The central goal of ranking is to exploit a relatively simple static measure to accurately select geological realizations that correspond to the targeted percentiles of the production responses, for example, those which represent P10, P50 and P90 (Ballin et al. 1992). This would define the bounds of the uncertainty without performing a large number of fine-scale flow simulations. The ranking and selection of realizations must be tailored to the flow process. It is well known that a particular ranking measure must be highly correlated to production response. Conventional ranking measures are, for example, original oil in place or connectivity (McLennan and Deutsch 2005). Use of streamline simulation (Gilman et al. 2002) and tracer simulation (Ballin et al. 1992) have received significant attention. However, there is no unique ranking index when there are multiple flow response variables and no ranking measure is perfect.

3.2.2 Comparison of Quantile Estimation

In this work, we have considered two different measures for each of the 405 realizations: original oil in place (OOIP) which represents the total volume of oil in the reservoir, and oil rate obtained by streamline tracer simulation, as used for the dissimilarity distance. Once a ranking measure is selected, the methodology for ranking and selecting the geostatistical realizations for flow processing is straightforward. The ranking measure is calculated for every geostatistical realization. The low (P10), medium (P50) and high (P90) geological realizations are then selected for flow modeling. Results for oil rate are presented in Fig. 11. We have presented the quantiles obtained with the entire set of realizations, and the quantiles resulting from the ranking measure. Figure 11(A) represents results using OOIP as a ranking measure. Figures 11(B) to 11(D) represent results using the oil rate from the tracer at respectively 7000, 10 000, and 20 000 days. Quantile estimations using ranking based on OOIP and streamline tracer are less accurate than the one obtained with the proposed methodology. To understand why the OOIP and tracer rankings provide less accurate results, we examine the correlation coefficients of the ranking measurement with the field oil rate. The correlation coefficient between OOIP and field oil rate is 0.19, which indicates that the OOIP is not a good measure for ranking. Tracer simulation is more suitable due to the improved correlation with the flow response (0.63, 0.76, and 0.87 at 7000, 10 000, and 20 000 days, respectively). This illustrates that in order for the ranking procedure to give reliable results, the ranking measure must be strongly correlated with production.

The ranking methods for selecting the P10, P50 and P90 realizations for flow simulation are not as accurate as the methodology proposed in this paper. However, only

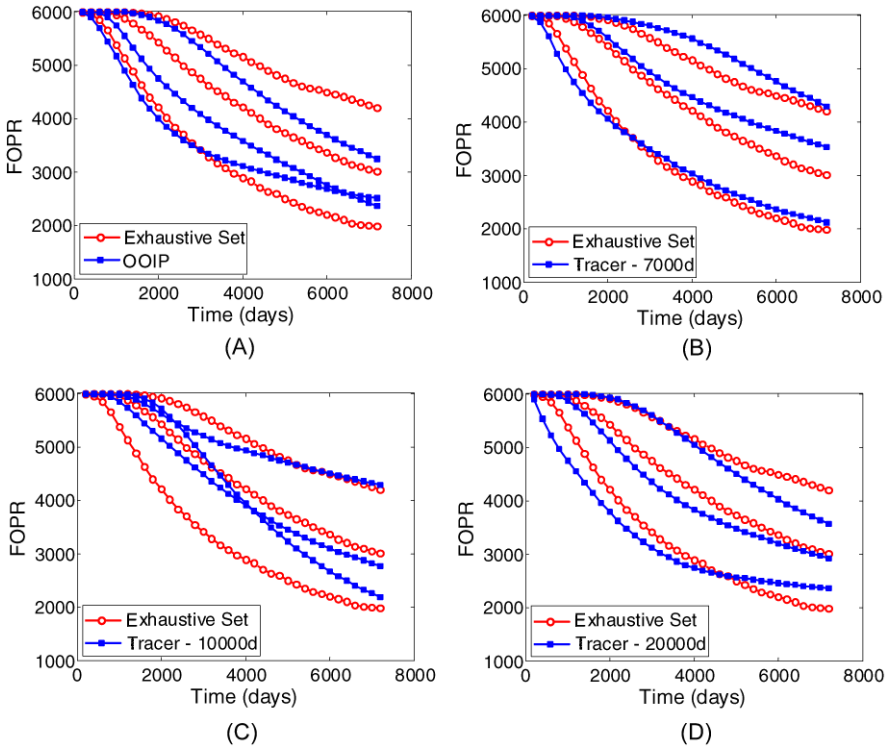


Fig. 11 Quantiles P10, P50 and P90 resulting from ranking measures: (**A**) OOIP, (**B**) Tracer—7000 days, (**C**) Tracer—10 000 days, (**D**) Tracer—20 000 days

3 full flow realizations were performed, whereas for the new method, 15 flow simulations were necessary. To compare both methods based upon the same number of flow simulations, we select 15 realizations equally spaced according to the ranking measure. Resulting quantiles for field oil rate are presented in Fig. 12. For the same number of flow simulation, Fig. 12 shows that the use of ranking measures is less accurate than the use of KPCA or KKM. Note that in the method proposed in this paper, we use tracer simulations for calculating the distance but not for ranking. The selection of realizations is done using another approach (KPCA or KKM). In the case shown here, results show that for the same measure (tracer simulation), better results are obtained from KPCA or KKM than from ranking (Fig. 13). In Fig. 13, we observe that the efficiency of the ranking technique relies on a high correlation coefficient between the ranking measurement and the flow response, whereas in case of distances, a smaller correlation coefficient is sufficient. In addition, for an equivalent correlation coefficient, quantiles are more accurate using distances and kernels, than using ranking measures (Fig. 14). In other words, the use of distances, instead of absolute values, improves the solution by using a "measure" to select the realizations.

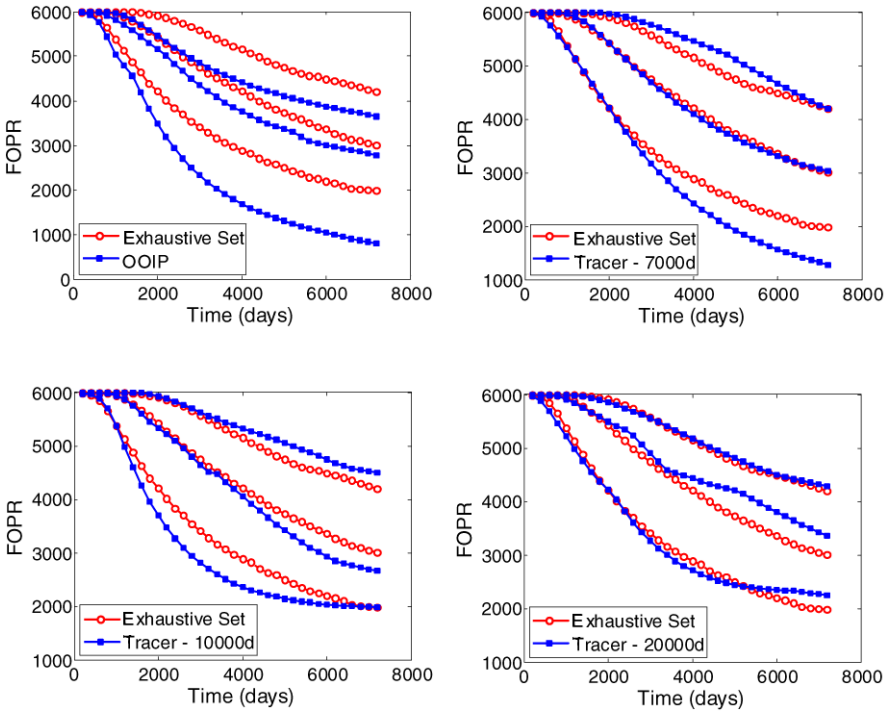


Fig. 12 Quantiles P10, P50 and P90 resulting from ranking measures: (**A**) OOIP, (**B**) Tracer—7000 days, (**C**) Tracer—10 000 days, (**D**) Tracer—20 000 days

4 Conclusions

We have presented results for a new realization-based method for uncertainty quantification of a spatial phenomenon. The method relies on a reasonable correlation between the distance measure and the response variables of interest. Using an application specific distance is an important additional tool which makes the task of response uncertainty quantification more effective. In general, each new type of application will require investigation of a new distance, which requires more work and produces more reward. For similar types of problems, these distances can then be reused. In our example of assessing subsurface flow uncertainty, we use streamline simulation to obtain the distances, which correlates well with the differences in production response using standard flow simulation. Given the distance measure, we employ KPCA and k-means clustering or kernel k-means to select a subset of 15 realizations which contains the same P10, P50, P90 quantiles as for the entire set of 405 models. The application of this new method shows promising results; quantile estimations using this methodology are noticeably better than those using traditional ranking methods for the same number of transfer function evaluations. In addition, only a small number of transfer function evaluations were necessary to obtain accurate uncertainty quantification though quantile estimation. An application to a simple synthetic case is provided in this paper. However, this methodology can also be successfully applied on oil and gas reservoirs, as presented in Scheidt and Caers (2008),

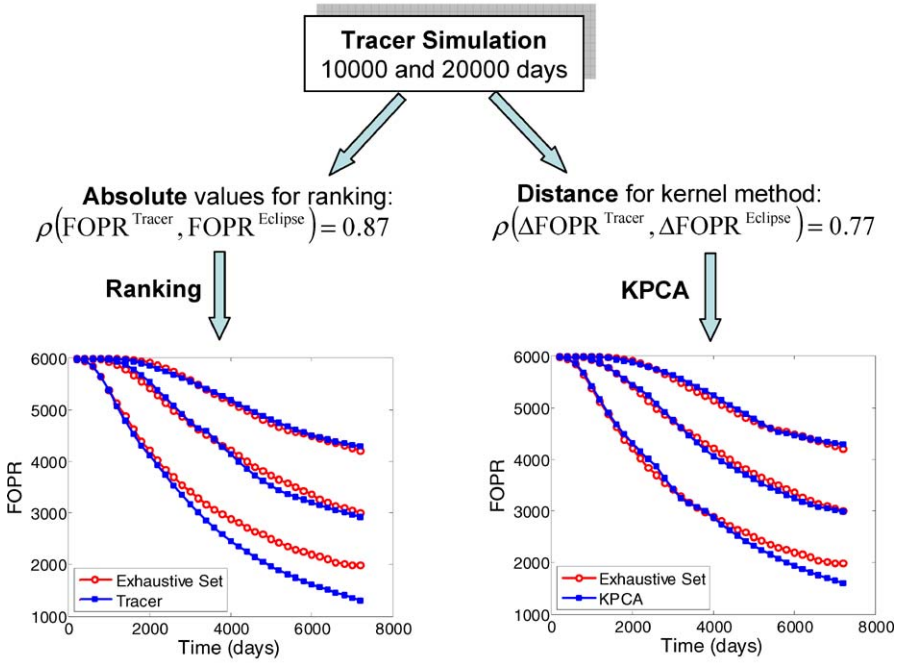


Fig. 13 Comparison between ranking and KKM using the same tracer measure. Note that the correlation coefficient for the absolute values of ranking measure is greater than the relative (distance) measure, but the P10, P50, P90 estimations are less accurate

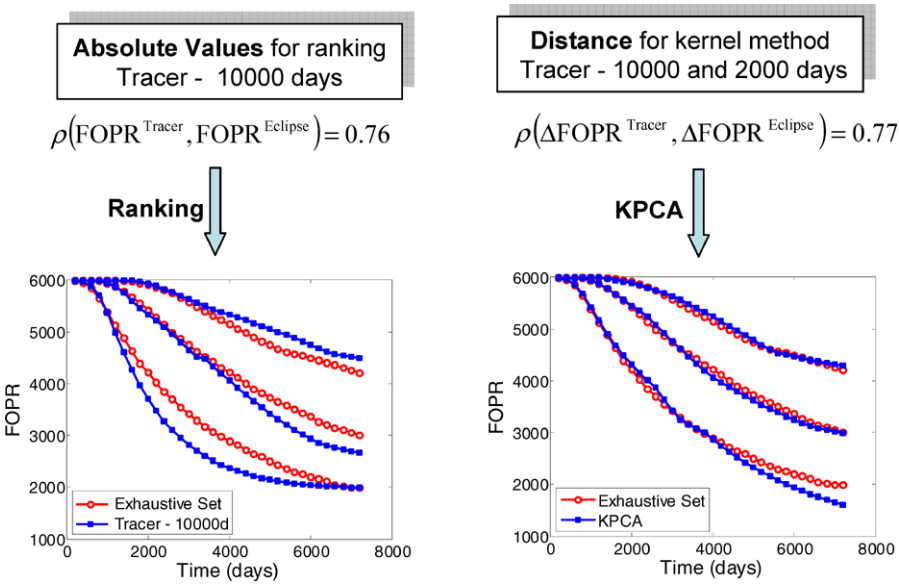


Fig. 14 Comparison between ranking and KKM for similar correlation coefficient. Note that for similar correlation coefficient, P10, P50, P90 estimations are more accurate for kernels

where it was shown that KPCA and KKM easily outperform the state of the art ranking technique. Tests on the robustness of the method with regards to the correlation between the distance and the difference of transfer function evaluations were performed on a real oil field case (Scheidt and Caers 2008). Results show that the higher the correlation, the smaller the error in the quantile estimation. In addition, if the correlation is low, increasing the number of transfer function evaluations is required in order to obtain accurate representation of uncertainty. However, no systematic bias or reduction in variance was noted. In the case where the correlation is zero, the method does not better or worse than random selection.

Acknowledgements The authors would like to acknowledge the SCRF sponsors and Chevron for their support. Many thanks as well to Darryl Fenwick from StreamSim Technologies for his help using 3DSL, and for many useful discussions.

References

Arpat BG (2005) Sequential Simulation with patterns. PhD dissertation, Stanford University, USA, 166 p
Arpat BG, Caers J (2007) Stochastic simulation with patterns. Math Geol 39(202):177–203
Ballin PR, Journel AG, Aziz K (1992) Prediction of uncertainty in reservoir performance forecast. JCPT 31(4):52–62
Batycky RP, Blunt MJ, Thiele MR (1997) A 3D field-scale streamline-based reservoir simulator. SPERE 12(4):246–254
Borg I, Groenen P (1997) Modern multidimensional scaling: theory and applications. Springer, New York, 614 p
Bowman AW, Azzalini A (1997) Applied smoothing techniques for data analysis. Oxford University Press, London, 204 p
Buhmann JM (1995) Data clustering and learning. In: Arbib M (ed) The handbook of brain theory and neural networks. MIT Press, Cambridge, pp 278–281
Box GEP, Draper NR (1975) Robust designs. Biometrica 62(2):347–352
Cox TF, Cox MAA (1994) Multidimensional scaling. Chapman and Hall, London, 213 p
Damslet E, Hage A, Volder R (1992) Maximum information at minimum cost! A north field development study with an experimental design. JPT 44(12):1350–1356
Dubuisson M-P, Jain AK (1994) A modified Hausdorff distance for object matching. In: Proceeding of the international conference on pattern recognition, Jerusalem, vol A, pp 566–568
Gilman JR, Meng H-Z, Uland MJ, Dzurman PJ, Cosic S (2002) Statistical ranking of stochastic geomodels using streamline simulation: a field application. In: SPE annual technical conference and exhibition. SPE 77374
Journel A, Alabert F (1990) New method for reservoir mapping. JPT 42(2):212–218
Keerthi SS, Lin C-J (2003) Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput 15(7):1667–1689
Manceau E, Zabalza-Mezghani I, Roggero F (2001) Use of experimental design to make decisions in an uncertain reservoir environment—from reservoir uncertainties to economic risk analysis. In: OAPEC conference, Rueil, France, 26–28 June 2001
McLennan JA, Deutsch CV (2005) Ranking geostatistical realizations by measures of connectivity. In: SPE/PS-CIM/CHOA, international thermal operations and heavy oil symposium, Calgary, Canada. SPE 98168-MS
Park K, Caers J (2007) History matching in low-dimensional connectivity-vector space. Unpublished SCRF report 20, Stanford University
Sarma P (2006) Efficient closed-loop optimal control of petroleum reservoirs under uncertainty. PhD dissertation, Stanford University, USA, 201 p
Scheidt C, Caers J (2008) A new method for uncertainty quantification using distances and kernel methods—application to a deepwater turbidite reservoir. SPE J (submitted)
Schöelkopf B, Smola A (2002) Learning with kernels. MIT Press, Cambridge, 664 p

Schöelkopf B, Smola A, Muller K-R (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1318

Shawe-Taylor J, Cristianni N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge, 462 p

Shi J, Malik J (2000) Normalized-cut and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905

Suzuki S, Caers J (2008) A distance based prior model parameterization for constraining solution of spatial inverse problems. Math Geosci 40(4):445–469

Vapnick VN (1998) Statistical learning theory. Wiley, New York, 736 p