Stochastic Environmental Research and Risk Assessment (2021) 35:457–480
https://doi.org/10.1007/s00477-020-01925-7

ORIGINAL PAPER

Conditional simulation of categorical spatial variables using Gibbs sampling of a truncated multivariate normal distribution subject to linear inequality constraints

Francky Fouedjio[1] · Celine Scheidt[2] · Liang Yang[1] · Yizheng Wang[1] · Jef Caers[1]

Accepted: 23 October 2020 / Published online: 6 November 2020

Abstract
This paper introduces a method to generate conditional categorical simulations, given an ensemble of partially conditioned (or unconditional) categorical simulations derived from any simulation process. The proposed conditioning method relies on implicit functions (signed distance functions) for representing the categorical spatial variable of interest. Thus, the conditioning problem is reformulated in terms of signed distance functions. The proposed approach combines aspects of principal component analysis and Gibbs sampling to achieve the conditioning of the unconditional categorical realizations to the data. It is applied to synthetic and real-world datasets and compared to the traditional sequential indicator simulation. It appears that the proposed simulation technique is an effective method to generate conditional categorical simulations from a set of unconditional categorical simulations.

Keywords Categorical spatial variable · Conditional simulation · Gibbs sampler · Implicit function · Principal component analysis

1 Introduction

Conditional simulations of categorical spatial variables in geostatistics are used to quantify spatial uncertainty relevant to variety of applications, such as environmental, groundwater, mineral, and oil/gas (Mariethoz and Caers 2014; Armstrong et al. 2011; Chiles and Delfiner 2012; Lantuejoul 2002; Deutsch 2002; Goovaerts 1998). Methods can be pixel-based, object-based or surface-based, although eventually all the results are rastered on a discrete mesh. In terms of pixel-based (or mesh-based) methods, one has variogram-based methods (Journel 1983; Deutsch 2006; Emery 2007), Markov-random field methods (Li 2007; Daly 2005; Elfeki and Dekking 2001; Tjelmeland and Besag 1998), and multiple-point-geostatistics methods (Strebelle 2002; Zhang et al. 2006; Arpat and Caers 2007; Mariethoz et al. 2010; Honarkhah and Caers 2010). Other works that deal with the problem of simulation of categorical spatial variables include methods based on spin models and maximum entropy (Žukovič and Hristopulos 2009; Bogaert and Gengler 2018). Conditioning to exact observations (hard data) is often easily achieved, simply because the central value to be simulated on a mesh is taken to be the same support as the hard data. Conditioning in object-based or surface-based methods is more challenging, since an object or surface is represented as a shape or geometry rather than a regular mesh. Additionally, object-based methods are more sensitive to any inconsistency between the hard data and the a-priori choice of model parameterization. Example of surface-based methods are those that involve physical processes in addition to a stochastic component. Physical processes such as sediment transport or geomorphology create surfaces, not regular meshes. Representing and conditioning these surfaces to exact observations remains challenging.

Object-based, surface-based, and process-based models often cannot be fully conditioned to hard data, in particular dense hard data; conditioning remains partial. In this paper,

✉ Francky Fouedjio
francky.fouedjio@stanford.edu

[1] Department of Geological Sciences, Stanford University, 367 Panama Street, Stanford, CA 94305, USA

[2] Department of Energy Resources Engineering, Stanford University, 367 Panama Street, Stanford, CA 94305, USA

we leverage on an ensemble of partially conditioned (or unconditional) categorical realizations generated by this kind of approaches and develop a method that creates conditional categorical simulations, without applying the original simulation method. Our idea relies on implicit functions. Implicit functions, sometimes also termed level set functions, represent surfaces (boundaries) by means of an additional dimension. For example, signed distance functions model the distance to a surface (boundary), adjusted by a sign to indicate inside/outside or above/below. This means that implicit functions transform and parameterize a 2D surface into a 3D mesh. We then use principal component analysis to create an orthogonal parameter representation of the implicit unconditional realizations. We show how the conditioning problem can be formulated as an inequality problem on the principal component scores. To then sample conditional realizations, we first sample principal component score from a multivariate Gaussian distribution with linear inequality constraints. The latter can also be seen as a truncated multivariate Gaussian distribution subject to linear inequalities constraints. Since the principal component orthogonalization and the signed distance function transformations are bijective, we can easily reconstruct conditional simulation, without applying the original simulation method.

The remainder of the paper is structured in the following manner. Sect. 2 details the proposed conditional simulation method through its basic ingredients. Section 3 analyzes the method on simple synthetic cases; in particular, the trade-off between the number of hard data and the number of unconditional categorical simulations is studied. Section 4 presents a real application in 3D to showcase the effectiveness of the proposed approach. A comparison with the classical sequential indicator simulation (SIS) is carried out. Section 5 outlines concluding remarks and suggestions for future work.

2 Methodology

Let $\{C(\mathbf{x}), \mathbf{x} \in D\}$ be the categorical spatial variable of interest defined on a fixed continuous spatial domain of interest $D \subset \mathbb{R}^p$, with a finite set of possible categorical outputs (categories) $\{c_1, \ldots, c_K\}$ which are mutually exclusive and collectively exhaustive. The categorical spatial variable of interest is observed at a set of n distinct locations $\{\mathbf{x}_i \in D\}_{i=1,\ldots,n}$. Consider an ensemble of L partially conditioned or unconditional categorical realizations $\{C_U^{(l)}(\mathbf{x}), \mathbf{x} \in D\}_{l=1,\ldots,L}$. Given this latter, the goal is to generate categorical realizations $\{C_C^{(t)}(\mathbf{x}), \mathbf{x} \in D\}_{t=1,\ldots,T}$ that honor the data $\{C(\mathbf{x}_i)\}_{i=1,\ldots,n}$, i.e., $C_C^{(t)}(\mathbf{x}_i) = C(\mathbf{x}_i),\ i = 1, \ldots, n$. This section is devoted to the description of the different ingredients required to implement the proposed conditioning method. This latter has been implemented in R environment (R Core Team 2020). The simulation of unconditional categorical realizations is not the concern here.

2.1 Implicit functions

Consider a bounded domain Ω in an area of interest $D \subset \mathbb{R}^p$ as illustrated on Fig. 1 (Zhou et al. 2016). Any point $\mathbf{x} \in D$ can obviously be classified into three parts: the interior or inside (Ω^-), the exterior or outside (Ω^+), and the boundary or surface ($\partial\Omega$). The $(p-1)$ dimensional boundary $\partial\Omega$ can be represented as the zero isocontour of a scalar function $\phi(\cdot)$ in $\mathbb{R}^p$, called implicit function or level set function: $\partial\Omega = \{\mathbf{x} \in D, \phi(\mathbf{x}) = 0\}$. The implicit function $\phi(\cdot)$ defines the boundary $\partial\Omega$ as well as the regions Ω^- and Ω^+ : $\Omega^- = \{\mathbf{x} \in D, \phi(\mathbf{x}) < 0\}$ and $\Omega^+ = \{\mathbf{x} \in D, \phi(\mathbf{x}) > 0\}$.

The signed distance function (SDF) is a subclass of implicit functions defined as (Osher and Fedkiw 2002):

$$\varphi(\mathbf{x}) = \begin{cases} -d(\mathbf{x}), & \text{if } \mathbf{x} \in \Omega^- \\ 0, & \text{if } \mathbf{x} \in \partial\Omega \\ d(\mathbf{x}), & \text{if } \mathbf{x} \in \Omega^+ \end{cases}, \tag{1}$$

where $d(\mathbf{x})$ denotes the minimum distance of $\mathbf{x}$ to $\partial\Omega$: $d(\mathbf{x}) = \min_{\mathbf{y} \in \partial\Omega} \|\mathbf{x} - \mathbf{y}\|$.

Thus, $\varphi(\cdot)$ takes on the distance from the boundary $\partial\Omega$ with a sign depending on being inside or outside the domain Ω. Signed distance functions are a subset of implicit functions having the property of unit gradient module with $\|\nabla d(\cdot)\| = 1$. Geometrically, it means that the δ isocontour of the signed distance function is the offset of its zero isocontour along the normal direction and the offset distance equals δ. Figure 2 shows two examples of boundaries and their corresponding signed distance functions on a 2-dimensional grid. Signed distance values along the boundary is equal to zero. Signed distance values inside the boundary are negative while signed distance values outside the boundary are positive.

The categorical spatial variable of interest $\{C(\mathbf{x}), \mathbf{x} \in D\}$ with K categories $\{c_1, \ldots, c_K\}$ can also be seen as a variable that creates distinct boundaries or surfaces in the spatial domain D. Implicit functions such as the signed distance functions are used to describe the categories. Each category $c_k (k = 1, \ldots, K)$ is represented by a signed distance function $\varphi_k(\cdot)$ such that: $c_k = \{\mathbf{x} \in D, \varphi_k(\mathbf{x}) < 0\}$ and its boundary $\partial c_k = \{\mathbf{x} \in D, \varphi_k(\mathbf{x}) = 0\}$. This means that any categorical realization, whether pixel-based, object-based or surface-based can be parameterized using signed distance functions. A categorical hard datum at

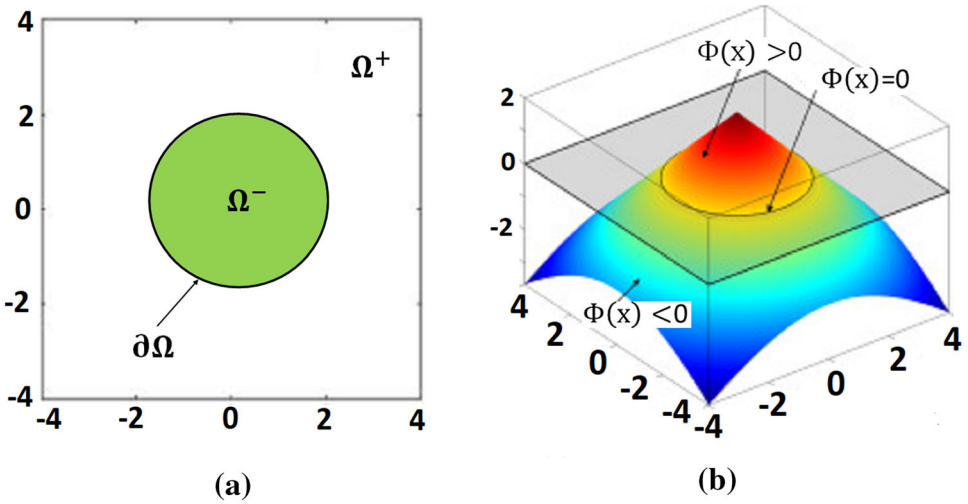


Fig. 1 Representation of a boundary $\partial\Omega$ as well as regions Ω^- and Ω^+ by means of an implicit function $\phi(\cdot)$

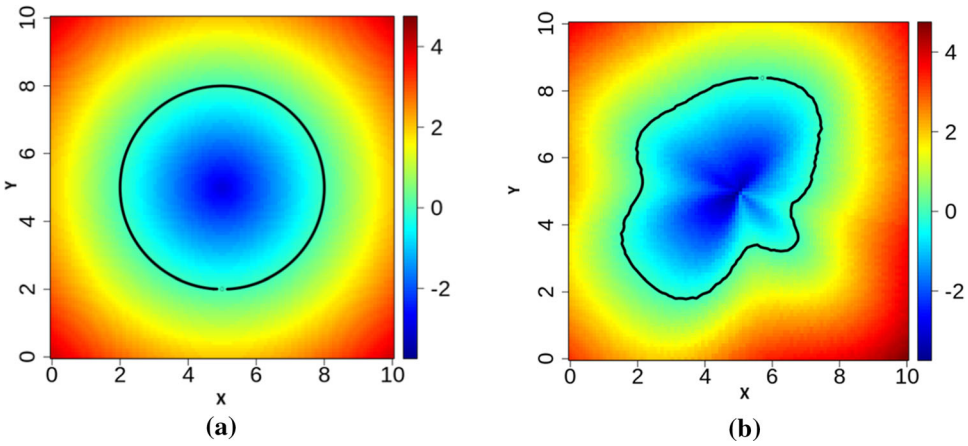


Fig. 2 Examples of boundaries (in black) and their associated signed distance functions (map) $\varphi(\cdot)$

location $\mathbf{x}$ then indicates the sign of all signed distance functions at $\mathbf{x}$. Additionally, the indicator notation is used to denote the presence or absence of a category c_k at a location $\mathbf{x} \in D$:

$$I_k(\mathbf{x}) = \begin{cases} 1, & \text{if } C(\mathbf{x}) = c_k \\ 0, & \text{if } C(\mathbf{x}) \neq c_k \end{cases}, \text{ and } I_k(\mathbf{x}) = 1 \text{ implies that } I_{k'}(\mathbf{x}) = 0 \quad \forall k' \neq k; k = 1, \ldots, K. \tag{2}$$

The categorical spatial variable of interest $\{C(\mathbf{x}), \mathbf{x} \in D\}$ with K categories can be transformed into a set of K signed distance functions $\{\varphi_k(\mathbf{x}), \mathbf{x} \in D\}_{k=1,\ldots,K}$ using the signed distance transform approach (Grevera 2007; Davies 2012). Each category c_k defines a p-dimensional binary image $\{I_k(\mathbf{x}), \mathbf{x} \in D\}$ where each point (pixel) has either a value of 1 indicating the presence of the category c_k or a value of 0 indicating the absence of the category c_k. For every point (pixel) set to 1, a distance transform assigns a value indicating the negatively signed distance from that point (pixel) to the nearest point (pixel) set to 0. Similarly for every point (pixel) set to 0, a distance transform assigns a value indicating the positively signed distance from that point (pixel) to the nearest point (pixel) set to 1. Additionally, the signed distance transformation is one-to-one. The bijectivity is obtained using the following rule:

$$I_k(\mathbf{x}) = \begin{cases} 1, & \text{if } \varphi_k(\mathbf{x}) = \min(\varphi_1(\mathbf{x}), \ldots, \varphi_K(\mathbf{x})) \\ 0, & \text{otherwise} \end{cases}, \quad \forall \mathbf{x} \in D,\ k = 1, \ldots, K. \tag{3}$$

Equivalent to

$$C(\mathbf{x}) = \underset{c_1,\ldots,c_k}{\arg\min} \left(\varphi_1(\mathbf{x}), \ldots, \varphi_K(\mathbf{x})\right), \ \forall \mathbf{x} \in D. \tag{4}$$

2.2 Principal component analysis (PCA)

By applying the signed distance transform approach described in Sect. 2.1, the ensemble of L unconditional categorical realizations $\{C_U^{(l)}(\mathbf{x}), \mathbf{x} \in D\}_{l=1,\ldots,L}$ is transformed into an ensemble of L unconditional signed distance realizations $\left\{\varphi_{U_k}^{(l)}(\mathbf{x}), \mathbf{x} \in D\right\}_{l=1,\ldots,L}$ for each category $c_k (k = 1, \ldots, K)$. Principal component analysis (PCA) performed on each ensemble $\left\{\varphi_{U_k}^{(l)}(\mathbf{x}), \mathbf{x} \in D\right\}_{l=1,\ldots,L}$ results in the following decomposition in finite dimensions:

$$\varphi_{U_k}^{(l)}(\mathbf{x}) = \sum_{k'=1}^{L} \alpha_{U_{k'},k}^{(l)} \psi_{U_{k'},k}(\mathbf{x}), \ \forall \mathbf{x} \in D, \ k = 1, \ldots, K; \qquad (5)$$

where $\{\alpha_{U_{k'},k}^{(l)}\}_{k'=1,\ldots,L}$ are principal component scores (coefficients) and $\{\psi_{U_{k'},k}(\mathbf{x}), \mathbf{x} \in D\}_{k'=1,\ldots,L}$ are principal components factors (basis functions).

In practice, the spatial domain D is represented with a number of grid cells (N). For each category $c_k (k = 1, \ldots, K)$, PCA is applied to a matrix $\boldsymbol{\Gamma}_{U_k}(L \times N)$ arranged as a set of L row vectors, each representing a single signed distance realization $\left\{\varphi_{U_k}^{(l)}(\mathbf{x}), \mathbf{x} \in D\right\}$. PCA can be performed in parallel for each matrix $\boldsymbol{\Gamma}_{U_k} (k = 1, \ldots, K)$. In Eq. (5), the number of principal components is equal to the number of realizations L. Indeed, in spatial problems, the size of the grid (number of grid cells) is usually much higher than the ensemble size used for quantifying spatial uncertainty ($L << N$).

In Eq. (5), the ensemble $\left\{\varphi_{U_k}^{(l)}(\mathbf{x}), \mathbf{x} \in D\right\}_{l=1,\ldots,L}$ can be seen as a set of images and $\left\{\varphi_{U_k}^{(l)}(\mathbf{x}), \mathbf{x} \in D\right\}$ as an image. Thus, the resulting principal components factors $\{\psi_{U_{k'},k}(\mathbf{x}), \mathbf{x} \in D\}_{k'=1,\ldots,L}$ are images as well. Hence, Eq. (5) provides a decomposition of the images into a set of eigen-images and a set of coefficients. Note that in PCA, the basis functions are fixed, while the coefficients are varying/random. A very useful property of PCA is the bijectivity. It allows the reconstruction of signed distances from coefficients. Moreover, the signed distance transformation is bijective as well. This means that a single set of principal component scores corresponds to a single and unique categorical realization.

The bijective nature of PCA means that an image can be reconstructed back, once all the principal component factors and scores are used. It is important to highlight that the PCA is used here more as an orthogonal decomposition approach than a dimension reduction technique because all the principal component factors are kept. The main reason for using principal component analysis based on the signed distance function instead of the random function modeling the categorical spatial variable of interest is that principal component analysis of categorical data is not very meaningful and in fact inefficient. The reason lies in the fact that a categorical realization contains less information than a signed distance function realization. The distance to a boundary is the additional information, not present in the categorical model.

2.3 Conditioning by the Gibbs sampler

Given the principal component factors $\{\psi_{U_{k'},k}(\mathbf{x}), \mathbf{x} \in D\}_{k'=1,\ldots,L}$ derived from the PCA decomposition of unconditional signed distance realizations as depicted by Eq. (5), the basic idea is to generate new principal component scores such that signed distance realizations match the data. Let

$$\varphi_k(\mathbf{x}) = \sum_{k'=1}^{L} \alpha_{k',k} \psi_{U_{k'},k}(\mathbf{x}), \quad \forall \mathbf{x} \in D, \quad k = 1, \ldots, K, \qquad (6)$$

where $\{\alpha_{k',k}\}_{k'=1,\ldots,L}$ are random coefficients. It is important to note that all the principal component factors are considered; so there is no truncation.

Principal component (PC) scores of signed distance realizations often show Gaussian type behavior (see Sect. 3.1). For each category $c_k (k = 1, \ldots, K)$, PC scores vector $\boldsymbol{\alpha}_k = (\alpha_{1,k}, \ldots, \alpha_{L,k})^T$ is assumed to follow a multivariate normal distribution defined by:

$$\boldsymbol{\alpha}_k \sim \exp\left(-\frac{1}{2}(\boldsymbol{\alpha}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\alpha}_k - \boldsymbol{\mu}_k)\right), \ k = 1, \ldots, K; \qquad (7)$$

where the mean $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$ are computed using unconditional PC scores $\{\alpha_{U_{k'},k}^{(l)}\}_{k'=1,\ldots,L}$ derived from the PCA decomposition of unconditional signed distance realizations given in Eq. (5). Specifically,

$$\begin{aligned} \boldsymbol{\mu}_k &= \left[\frac{1}{L}\sum_{l=1}^{L} \alpha_{U_{k'},k}^{(l)}\right]_{k'=1,\ldots,L}; \\ \boldsymbol{\Sigma}_k &= \frac{1}{L-1}\sum_{l=1}^{L} (\boldsymbol{\alpha}_{U_k}^{(l)} - \boldsymbol{\mu}_k)(\boldsymbol{\alpha}_{U_k}^{(l)} - \boldsymbol{\mu}_k)^T, \\ &\text{with } \boldsymbol{\alpha}_{U_k}^{(l)} = \left[\alpha_{U_{k'},k}^{(l)}\right]_{k'=1,\ldots,L}. \end{aligned} \qquad (8)$$

The covariance matrix $\boldsymbol{\Sigma}_k$ is a diagonal matrix because the PC scores are uncorrelated by construction. Thus, the normality assumption of PC scores can be checked using classical tools in the univariate setting (e.g., normal probability plot, quantile-quantile plot, Kolmogorov-Smirnov test, Shapiro-Wilk test).

As indicated above, hard data informs the sign of the signed distance function associated with a category. Hence, the set of hard data can be translated into as set of inequality constrains using Eq. (6). For example, assume that at the data location $\mathbf{x}_1$, the category c_2 is observed ($C(\mathbf{x}_1) = c_2$). This means that the signed distance function associated with the category c_2 should be negative at location $\mathbf{x}_1$ ($\varphi_2(\mathbf{x}_1) \leq 0$), and the signed distance functions associated with other categories should be positive at location $\mathbf{x}_1$ ($\varphi_k(\mathbf{x}_1) \geq 0,\ \forall k \neq 2; k = 1, \ldots, K$). For each $\varphi_k (k = 1, \ldots, K)$, the conditioning to all data locations is expressed by the following inequalities:

$$\begin{cases} \varphi_k(\mathbf{x}_1) = \alpha_{1,k}\psi_{U_{1,k}}(\mathbf{x}_1) + \alpha_{2,1}\psi_{U_{2,k}}(\mathbf{x}_1) + \ldots + \alpha_{L,k}\psi_{U_{L,1}}(\mathbf{x}_1) \leq 0 \text{ or } \geq 0 \\ \ldots \\ \varphi_k(\mathbf{x}_n) = \alpha_{1,k}\psi_{U_{1,k}}(\mathbf{x}_n) + \alpha_{2,k}\psi_{U_{2,k}}(\mathbf{x}_n) + \ldots + \alpha_{L,k}\psi_{U_{L,k}}(\mathbf{x}_n) \leq 0 \text{ or } \geq 0 \end{cases} \tag{9}$$

In Eq. (9), the n inequalities corresponding to n hard data can be summarized by:

$$\boldsymbol{\Psi}_{U_k}\boldsymbol{\alpha}_k \leq \mathbf{0},\ k = 1, \ldots, K. \tag{10}$$

In Eq. (7), since individual elements of $\boldsymbol{\alpha}_k = (\alpha_{1,k}, \ldots, \alpha_{L,k})^T$ belong to $\mathbb{R}$, one can write $-\infty \leq \alpha_{k',k} \leq +\infty,\ k' = 1, \ldots, L$. Thus, Eqs. (7) and (10) define a truncated multivariate normal (TMVN) distribution subject to linear inequalities. A good collection of statistical properties of TMVN distributions can be found in Horrace (2005). Thus, the exact conditioning problem is equivalent to the construction of samples from a TMVN distribution subject to linear inequality constraints. Sampling from such distribution proceeds through the Gibbs sampler (Li and Ghosh 2015). Gibbs sampling is a Markov Chain Monte Carlo (MCMC) algorithm where each random variable is iteratively resampled from its conditional distribution given the remaining variables. Gibbs sampling is a good candidate for this task as all full conditional distributions of a truncated multivariate normal distribution are truncated univariate normal (TUVN) distributions. Gibbs sampling requires initial values $\boldsymbol{\alpha}_k^{(0)} = (\alpha_{1,k}^{(0)}, \ldots, \alpha_{L,k}^{(0)})^T$ that already satisfy the inequality constraints given in Eq. (10). To do this, a simple linear programming is performed to find an initial solution to the system of inequality constraints (Vanderbei 2013).

Given the Gibbs samples $\{\boldsymbol{\alpha}_k^{(t)}\}_{t=1,\ldots,T}$, conditional categorical realizations are given by the following truncation rule:

$$C_C^{(t)}(\mathbf{x}) = \underset{c_1,\ldots,c_k}{\arg\min} \left(\varphi_{C_1}^{(t)}(\mathbf{x}), \ldots, \varphi_{C_K}^{(t)}(\mathbf{x})\right),\ \forall \mathbf{x} \in D; \tag{11}$$

where $\varphi_{C_k}^{(t)}(\mathbf{x}) = \sum_{k'=1}^{L} \alpha_{k',k}^{(t)}\psi_{U_{k',k}}(\mathbf{x})$ is a conditional signed distance realization associated with category $c_k (k = 1, \ldots, K)$.

If the categories $c_1, \ldots, c_K$ present a certain order (or sequence), the truncation rule defined in Eq. (11) can be modified to account this order as illustrated in Yang et al. (2019). It is important to highlight that the number of conditional categorical simulations T can be greater or less than the number of unconditional categorical simulations L. The Gibbs sampler here can be performed in parallel for each $\boldsymbol{\alpha}_k (k = 1, \ldots, K)$ as well as for each conditional categorical realization $t (t = 1, \ldots, T)$. As with other MCMC sampling methods, Gibbs sampler generates a Markov chain of samples, each of which is correlated with nearby samples. As a result, care must be taken to obtain independent samples. It is common to sample T draws and discard the first B, as burn-in, and then retain every sth sample.

Given the set of L unconditional categorical simulations $\{C_U^{(l)}(\mathbf{x}), \mathbf{x} \in D\}_{l=1,\ldots,L}$ and the data $\{C(\mathbf{x}_i)\}_{i=1,\ldots,n}$, the proposed conditional simulation method performs the following steps:

1. For $k = 1, \ldots, K$, generate the ensemble of L unconditional signed distance realizations $\left\{\varphi_{U_k}^{(l)}(\mathbf{x}), \mathbf{x} \in D\right\}_{l=1,\ldots,L}$ using the signed distance transform method;
2. For $k = 1, \ldots, K$, apply PCA to the matrix $\boldsymbol{\Gamma}_{U_k}$ arranged as a set of L row vectors, each representing a single unconditional signed distance realization $\left\{\varphi_{U_k}^{(l)}(\mathbf{x}), \mathbf{x} \in D\right\}$; the PCA outcomes are unconditional PC scores $\{\alpha_{U_{k',k}}^{(l)}\}_{k'=1,\ldots,L}$ and PC factors $\{\psi_{U_{k',k}}(\mathbf{x}), \mathbf{x} \in D\}_{k'=1,\ldots,L}$;
3. For $k = 1, \ldots, K$, generate conditional PC scores $\{(\alpha_{1,k}^{(t)}, \ldots, \alpha_{L,k}^{(t)})^T\}_{t=1,\ldots,T}$ using the Gibbs sampling approach;
4. For $k = 1, \ldots, K$, derive conditional signed distance realizations by reconstruction $\varphi_{C_k}^{(t)}(\mathbf{x}) = \sum_{k'=1}^{L} \alpha_{k',k}^{(t)}\psi_{U_{k',k}}(\mathbf{x})$;
5. Apply the truncation rule $C_C^{(t)}(\mathbf{x}) = \underset{c_1,\ldots,c_k}{\arg\min} \left(\varphi_{C_1}^{(t)}(\mathbf{x}), \ldots, \varphi_{C_K}^{(t)}(\mathbf{x})\right),\ \forall \mathbf{x} \in D$ for obtaining conditional categorical simulations.

2.4 Falsification of unconditional categorical simulations

Most of conditioning approaches, implicitly assume that unconditional simulations and data are consistent. In a

Bayesian sense, the prior distribution may not predict the data. Thus, before performing the conditioning of the unconditional categorical simulations to the data, it is important to test if these unconditional categorical simulations are consistent in Bayesian sense with the data. This is achieved by means of a falsification procedure (Scheidt et al. 2018). If unconditional categorical simulations are falsified, the resulting conditional simulations might not reproduce some statistical properties (e.g., mean and variance).

Let $\mathbf{d}^{(0)}$ be the vector of observed values at data locations (termed actual dataset) and $\{\mathbf{d}^{(l)}\}_{l=1,\ldots,L}$ be the vector of simulated values at data locations (termed simulated dataset). Unconditional categorical simulations are falsified if the actual dataset $\mathbf{d}^{(0)}$ is not within the same population as the simulated datasets $\{\mathbf{d}^{(l)}\}_{l=1,\ldots,L}$, i.e. $\mathbf{d}^{(0)}$ is an outlier. The idea consists in performing a multivariate outlier detection through a hypothesis test. This latter can not be performed directly on datasets $\{\mathbf{d}^{(l)}\}_{l=0,\ldots,L}$ because they are categorical information. However, pairwise distances among the datasets (including the actual dataset) can be calculated using a distance measure dedicated to categorical data. Here we choose the Jaccard distance which is one of most popular distance measures for categorical data; however, other distance measures can be used as well. Given the $(L+1)\times(L+1)$ distance matrix between datasets, the Multidimensional scaling (MDS) (Borg and Groenen 2007) can be applied on this latter to map the datasets $\{\mathbf{d}^{(l)}\}_{l=0,\ldots,L}$ into an Cartesian space such that distances between points in this space reflect the pairwise distances among the datasets. Thus, each dataset $\mathbf{d}^{(l)}(l=0,\ldots,L)$ (including the actual dataset) is represented by a m-dimensional point $\mathbf{y}_l$ $(m\leq(L+1))$ in the MDS space. To test, if unconditional categorical simulations and actual data are consistent, a statistical procedure based on robust Mahalanobis distance in the MDS space is used.

The robust Mahalanobis distance (RMD) for each dataset (including the actual dataset) is computed as follows:

$$RMD(\mathbf{d}^{(l)})=\sqrt{(\mathbf{y}_l-\hat{\mathbf{m}})\hat{\mathbf{C}}^{-1}(\mathbf{y}_l-\hat{\mathbf{m}})^T}\quad l=0,\ldots,L; \tag{12}$$

where $\mathbf{y}_l$ are the coordinates of the dataset $\mathbf{d}^{(l)}$ in the MDS space; where $\hat{\mathbf{m}}$ and $\hat{\mathbf{C}}$ are the robust estimation of mean and covariance of $\mathbf{Y}=[\mathbf{y}_0,\ldots,\mathbf{y}_{L+1}]^T$ (Hubert et al. 2018).

Assuming that $\mathbf{y}_l$ is multivariate normally distributed, Mahalanobis distance $[RMD(\mathbf{d}^{(l)})]^2$ follows a Chi-square distribution with m degrees of freedom χ^2_m. The 97.5 percentile of $\sqrt{\chi^2_m}$ is used as the cutoff. Thus, if the $RMD(\mathbf{d}^{(0)})$ falls outside the tolerance, i.e., $RMD(\mathbf{d}^{(0)})>\sqrt{\chi^2_{m,0.975}}$, the $\mathbf{d}^{(0)}$ is considered as an outlier, which means unconditional categorical simulations are not consistent with the actual observations, hence are falsified. Although multivariate outlier detection based on the Mahalanobis distance has the advantage of providing robust statistical calculations, it relies on the Gaussian assumption of marginal distribution of data. In the case where this assumption is doubtful, other multivariate outlier detection techniques such as one-class SVM (Schölkopf et al. 2001), Isolation Forest (Liu et al. 2008), local outlier factor (Breunig et al. 2000) can be used.

3 Simulation study

This section first illustrates the method with a simple case, then studies the various elements, components, and parameters.

3.1 Illustration of the method

Consider an object-based simulation, where two sources of information are available: the density of objects (Fig. 3a) and drill-holes with observations of absence (category 1) versus presence (category 2) (Fig. 3b, c). The total amount of hard data is $n=300$. Constraining object models to dense drill-hole data is challenging, while constraining to a density function is not. In that respect, we first generate an ensemble of 100 object realizations that are constrained to the density, but not to the drill-holes. Constraining to density is done through a rejection method (Lantuejoul 2002). Since there are only two categories, the signed distance function associated with one category is the opposite of the signed distance function associated with another category. So, it sufficient here to consider only one category. We will consider the category 2 (see, Fig. 3b, c).

Figure 4 shows three unconditional categorical realizations as well as the corresponding unconditional signed distance realizations for category 2. Figure 5 illustrates the consistency between unconditional categorical realizations and data. Figure 5a–c show the actual dataset $\mathbf{d}^{(0)}$ and the simulated datasets $\{\mathbf{d}^{(l)}\}_{l=1,\ldots,100}$ in the MDS space. The computed RMD for the actual dataset $RMD(\mathbf{d}^{(0)})$ and for the simulated datasets $\{RMD(\mathbf{d}^{(l)})\}_{l=1,\ldots,100}$ are given in Fig. 5d. The computed RMD for the actual dataset falls below the 97.5 percentile threshold which is equal to 6.2. Thus, unconditional categorical realizations are consistent with data.

To initialize the Gibbs sampling we perform first the linear programming to get an initial solution. Although

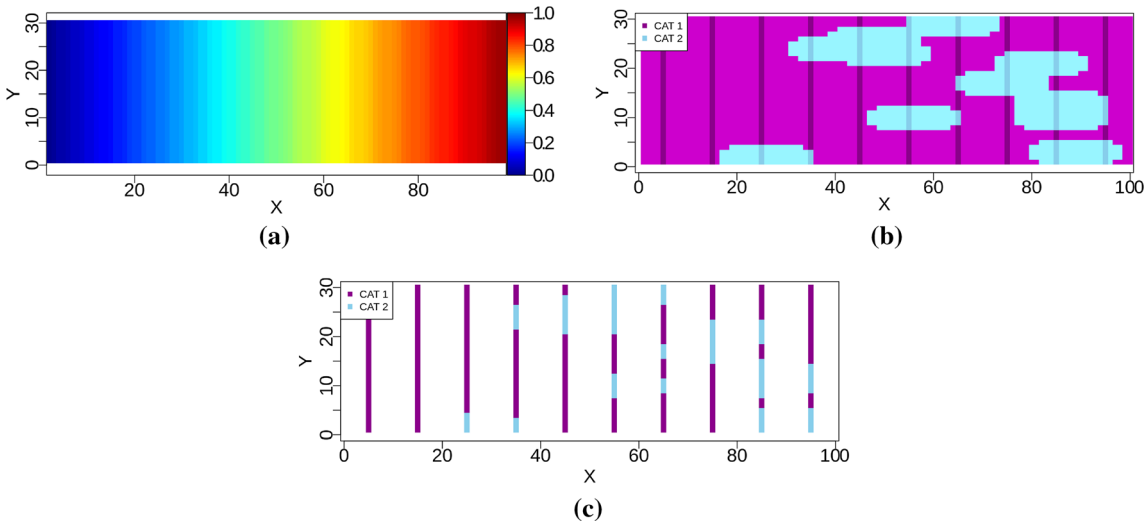


Fig. 3 **a** intensity map, **b** reference categorical model, **c** drill-hole data

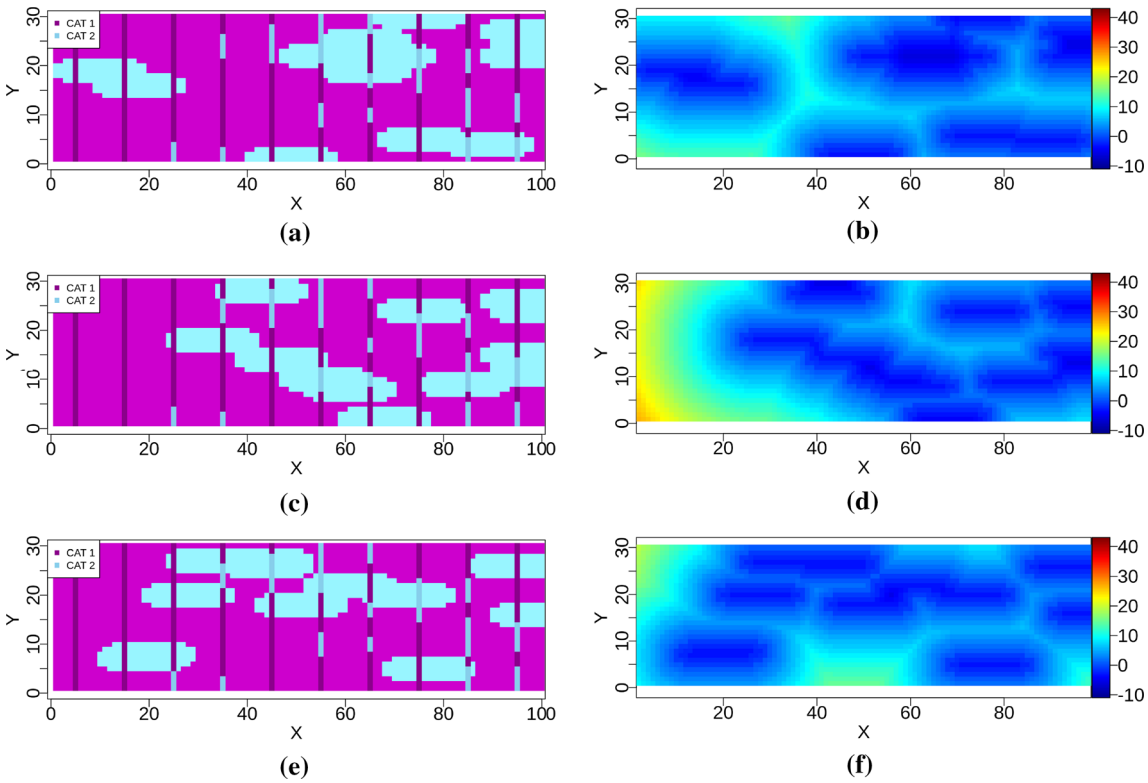


Fig. 4 **a**, **c**, **e** unconditional categorical realizations. **b**, **d**, **f** corresponding unconditional signed distance realizations for category 2

linear programming finds the optimum of a (linear) objective function subject to inequality constraints, here the objective function is simply a constant, hence it only finds a solution that follows the inequality constraints. Additionally, linear programming finds only one solution. Note that in this case there are 300 inequality constraints. Next, Gibbs sampling is performed with $T = 60{,}000$ iterations. It took approximately 30 minutes on a desktop computer (LINUX environment) with Intel(R) Core(TM) i9-7900X CPU @ 3.30GHz (10 cores / 20 threads), and 120 GB RAM.

Figures 6 and 7 show the trace plot and running average of the first four scores obtained by the Gibbs sampling. One can see how well the chain is mixing. Both the individual samples and the running average of the samples oscillate around a stable value of $\boldsymbol{\alpha}_k$, which is an indication of

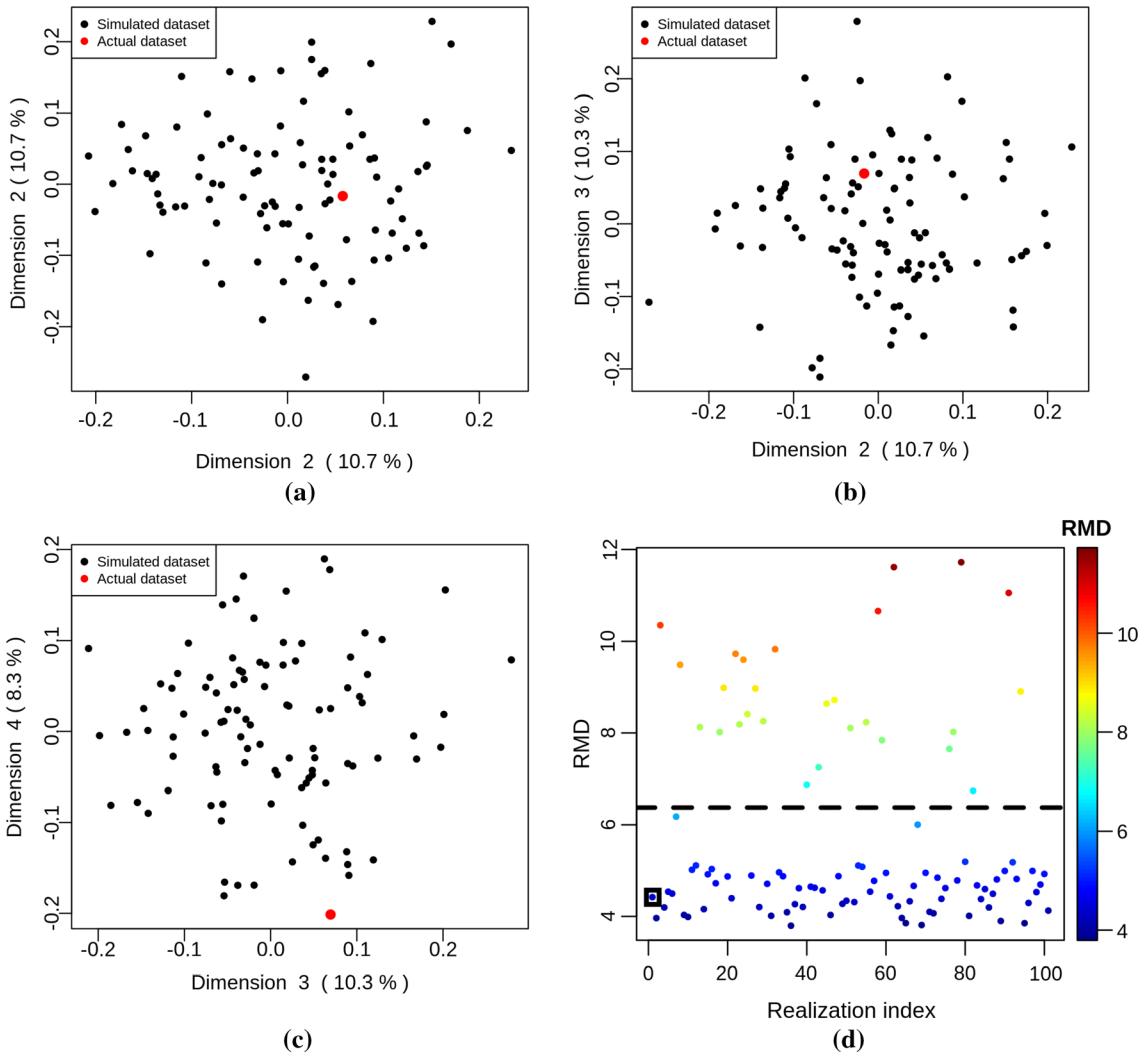


Fig. 5 Falsification of unconditional categorical simulations using robust Mahalanobis distance (RMD). **a**–**c** coordinates of datasets (actual and simulated) in the MDS space. **d** circle dots represent the calculated RMD for datasets (actual and simulated). The black-squared dot is the RMD for the actual dataset. The black dash line is the 97.5 percentile of the Chi-Squared distributed RMD

convergence of the chain. Typically, in Markov Chain Monte Carlo applications, initial samples are discarded to ensure that the Markov Chain has stabilized to the stationary distribution. This is referred to as burn-in samples. A burn-in period of $B = 10{,}000$ is considered. Another way to check for convergence is to analyze the autocorrelation between samples (Fig. 8). The lag-k autocorrelation is the correlation between every sample and the sample k steps before. As expected, the autocorrelation is high between consecutive samples, and decreases as the separation between samples increases. In this particular example, samples taken $s = 500$ samples apart can be considered as independent. Thus, there are 100 conditional categorical realizations retained. If the autocorrelation remains high for large values of k, this indicates a poor mixing of the chain. In this case, the Gibbs sampler should be performed from different initial solutions for improving mixing.

Figure 9 shows the unconditional and conditional PC scores for the first 4 PCs. We also apply a test for Gaussianity that shows that the unconditional PC scores follow a distribution that is close to Gaussian, see Fig. 10. Figure 11 shows some conditional categorical realizations. All conditional realizations match the data perfectly. The average proportion of objects for conditional categorical realizations is 26%, which is similar to the input proportion of the reference model (27%). We do notice some degradation in the object geometries. The conditional mean and variance maps are depicted by Fig. 12.

3.2 Monte Carlo evaluation of the method

The proposed conditioning approach relies on an ensemble of unconditional categorical realizations to performing the conditioning on the data. We want study the effect of the

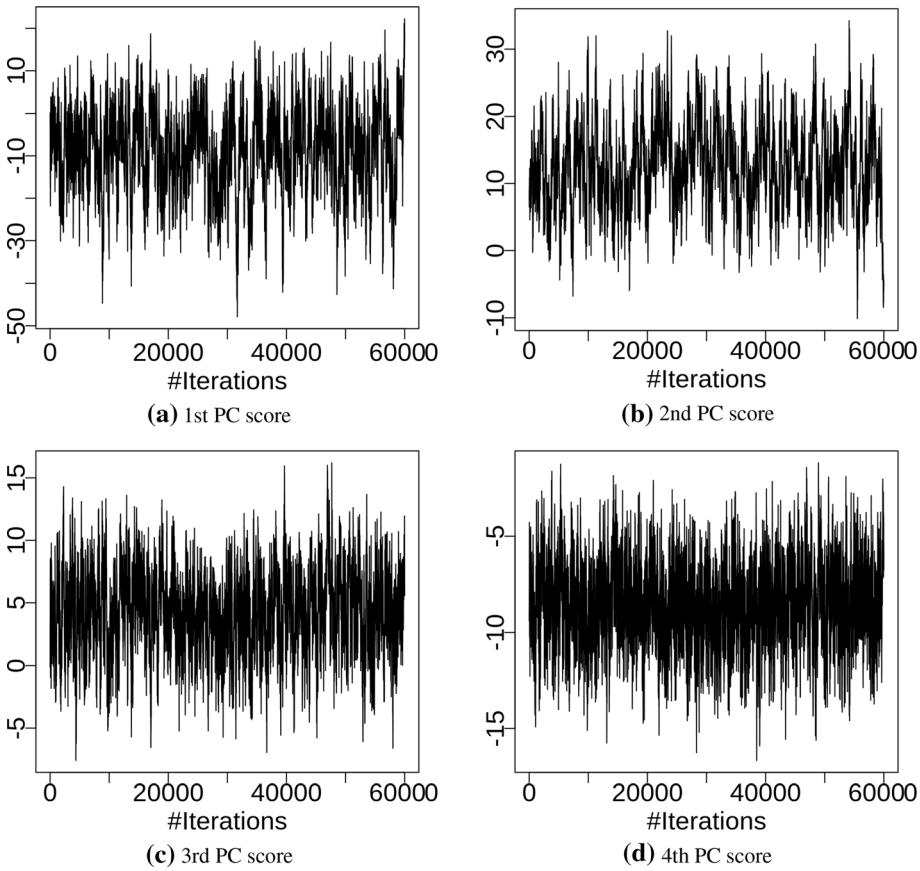


Fig. 6 Trace plot of the first 4 scores in the Gibbs sampler

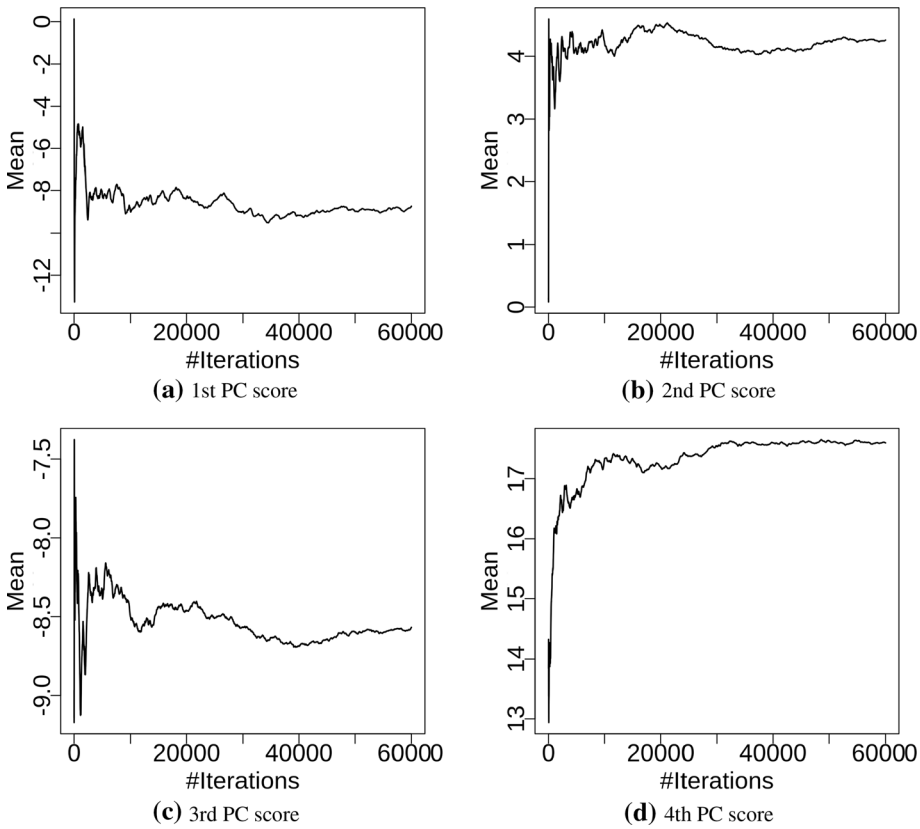


Fig. 7 Running average of the first 4 scores in the Gibbs sampler

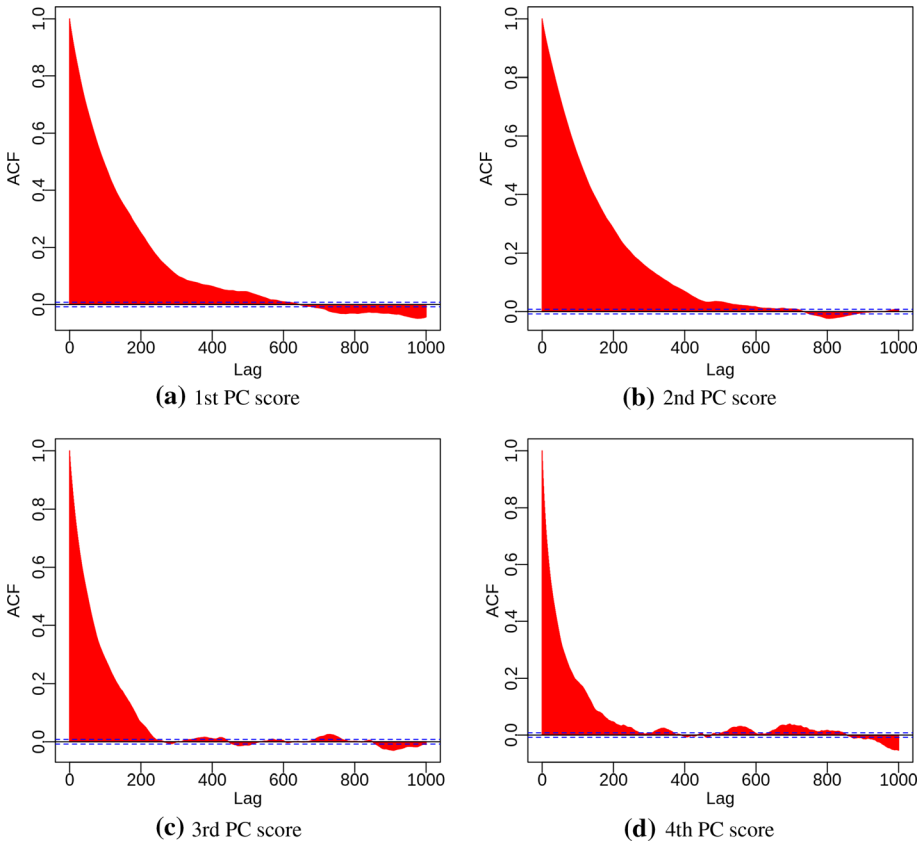


Fig. 8 Autocorrelation of the first 4 scores in the Gibbs sampler

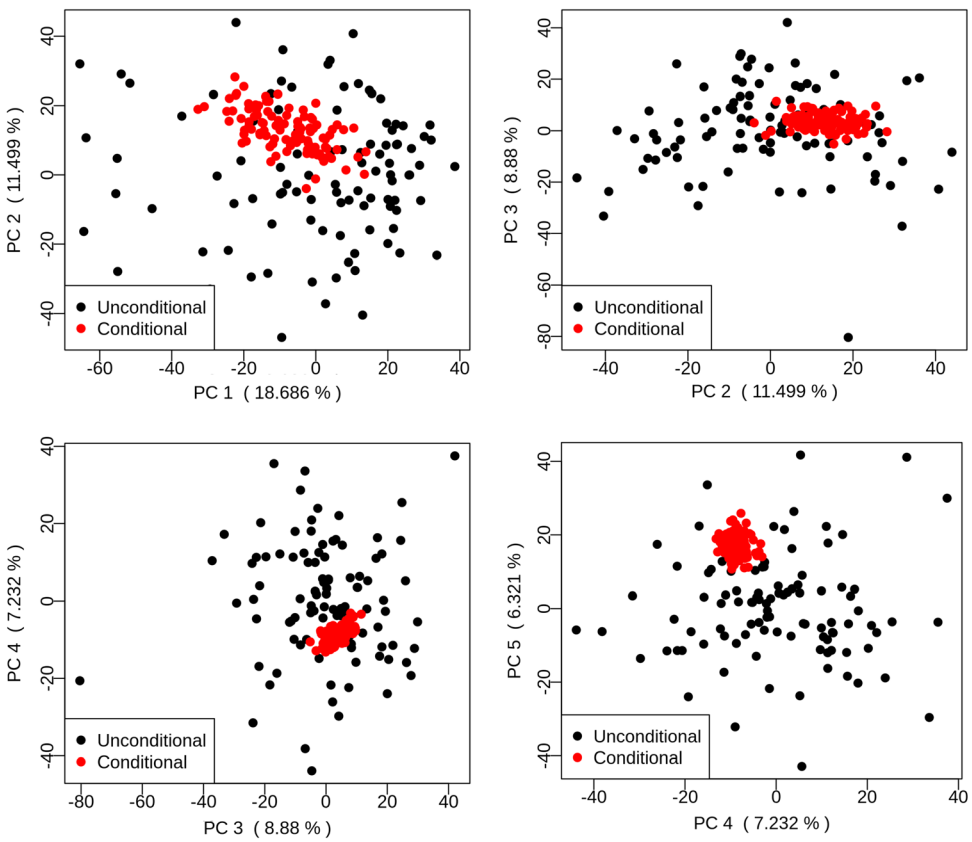


Fig. 9 100 unconditional first 4 PC scores and 100 conditional first 4 PC scores

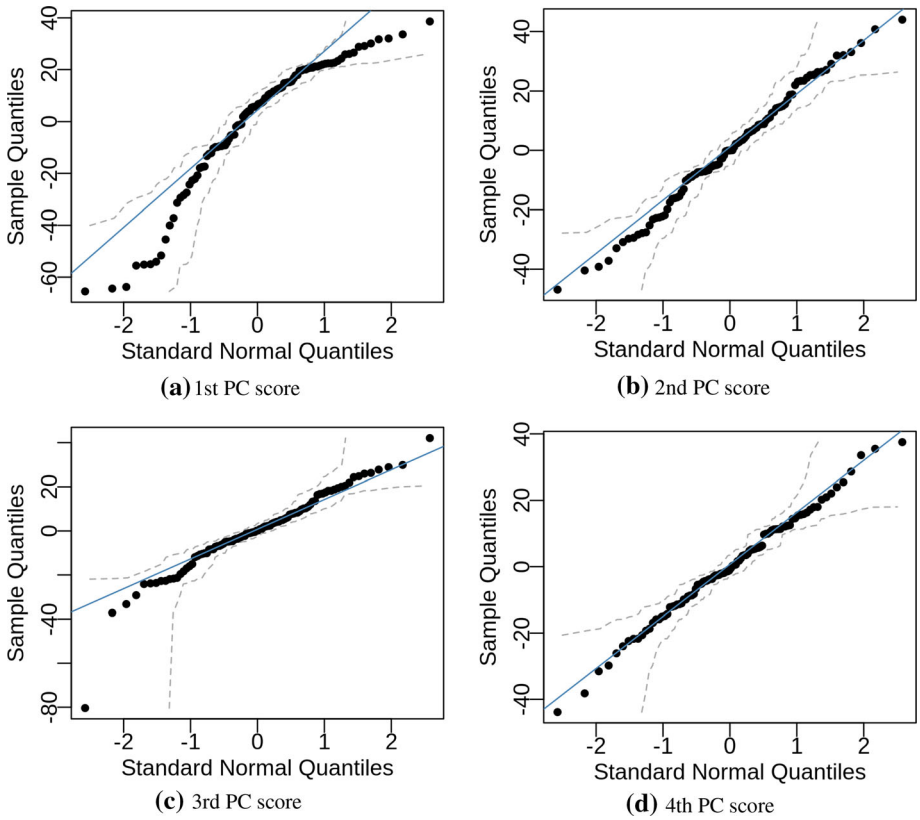


Fig. 10 QQ-plot of the first 4 scores compared to Gaussian distribution

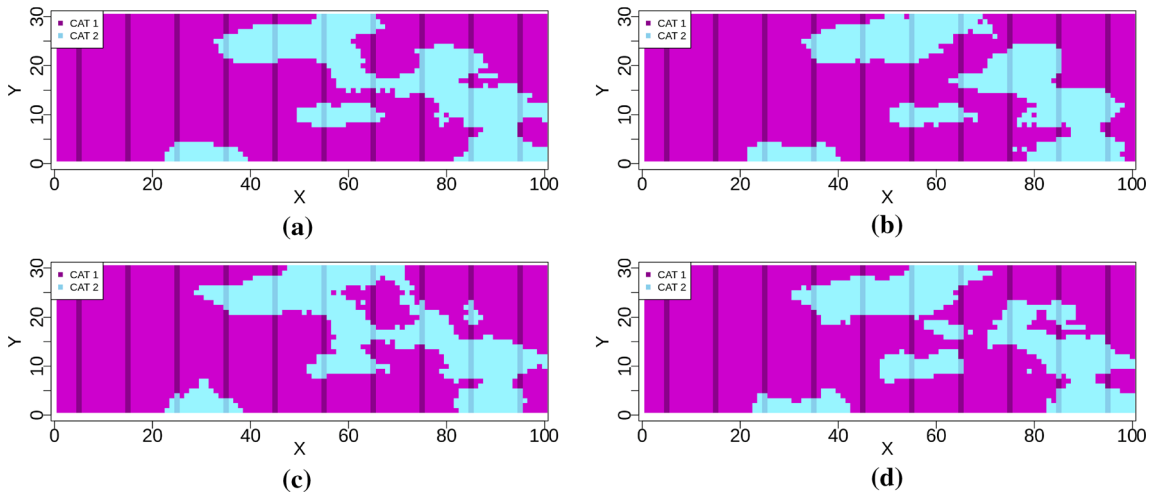


Fig. 11 4 out 100 conditional categorical simulations from the proposed conditioning method

number of unconditional categorical realizations versus the number of hard data points on the spatial uncertainty generated with the proposed method. Obviously, enough unconditional categorical realizations are needed to inform the statistical model in the a-priori variation of principal components. An additional condition is brought by the amount of hard data. Hard data are translated into inequality constraints. Too many constraints relative to too few unconditional categorical realizations will lead to too small uncertainty. Our study therefore requires some reference uncertainty. This reference uncertainty is simply a case with a very large amount of unconditional categorical realizations, here 1000. Then we study what happens when the number of unconditional categorical realizations is reduced. A third factor is the nature of the unconditional categorical model. A more spatially correlated unconditional categorical model would need less unconditional categorical realizations.

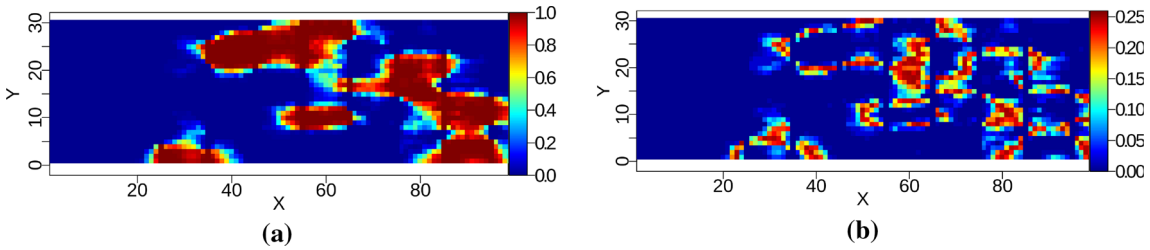


Fig. 12 **a** conditional mean and **b** conditional variance computed from 100 conditional categorical simulations generated using $L = 100$ unconditional categorical simulations

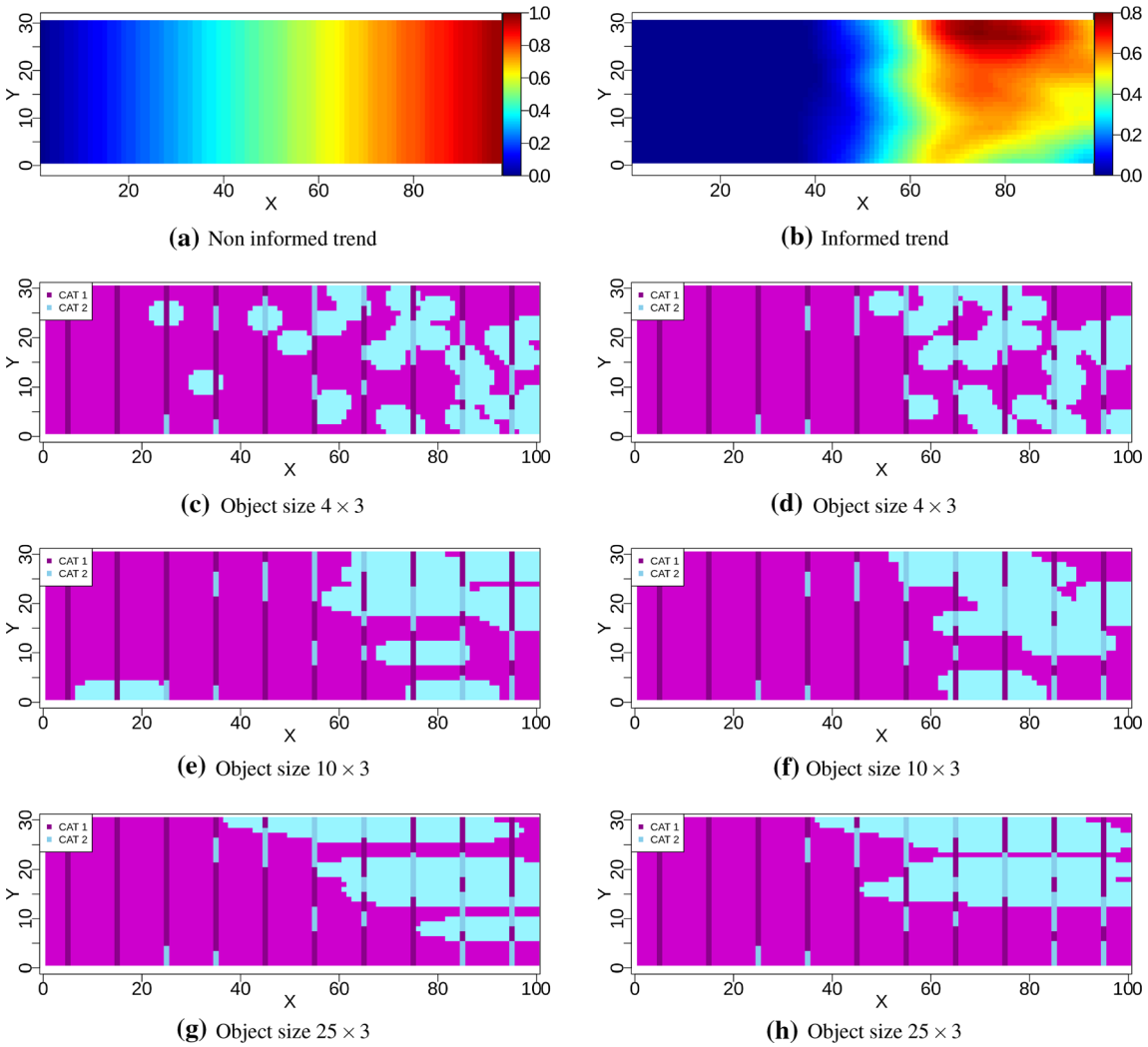


Fig. 13 Spatial model uses tree alternate object sizes and two intensity maps

Our Monte Carlo study therefore varies the number of unconditional categorical models but also varies the spatial model itself. To that extent, we use three alternate object sizes, see Fig. 13, one smaller, one medium and one larger; as well as two different intensity maps, one more constraining. The intent is to study the impact of spatial variability. Of note is the definition of the intrinsic dimensionality of the hard data. The extrinsic

Table 1 Intrinsic dimension H for each case

Ellipse size	Less informed trend	More informed trend
Small	44	31
Medium	27	22
Large	16	14

dimensionality is 300. However, because of spatial correlation, the intrinsic dimensionality is much less. One way to define this is to look at the variability of the unconditional signed distance realizations at the data locations. In other words, we assign, at each data location the unconditional signed distance realizations and perform the PCA. Then observe dimension H at the 95% variance cut-off. Table 1 shows that the intrinsic dimension H is much less than the number of hard data n and dependent on the unconditional categorical model.

To establish a measure of distance between the reference uncertainty model and the results, we use the following metric:

$$\epsilon_L = \frac{\sum_{i=1}^{N} |Var_i^L - Var_i^{1000}|}{N}, \tag{13}$$

which is the average absolute deviation in terms of the conditional ensemble variance. $N = 3000$ refers to the size of entire grid of 100×30. The index i in Eq. (13) denotes the grid node. Figures 14 and 15 show how this works for one particular case (less informed trend and large object size). In this case the intrinsic dimension is $H = 16$. We notice how the conditional variance starts to deviate significantly after 75 unconditional realizations, much less than the initial 300. Figure 16 (left) summarizes all the result. For each case, the deviation in variance decreases as the number of unconditional realizations increases. In addition, the deviation in variance is smaller for larger ellipses and for models with a more informed trend. Because of the linearity observed in Fig. 16 (right), we derive a rule of thumb as

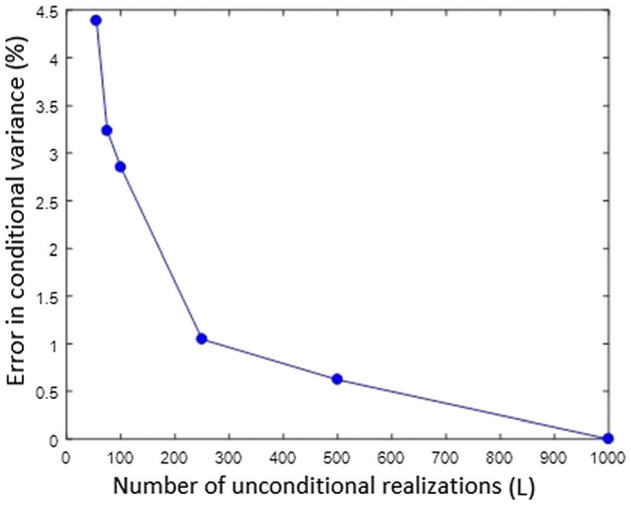


Fig. 15 Relationship between the number of unconditional categorical simulations and the error in conditional variance for the case with less informed trend and large object size

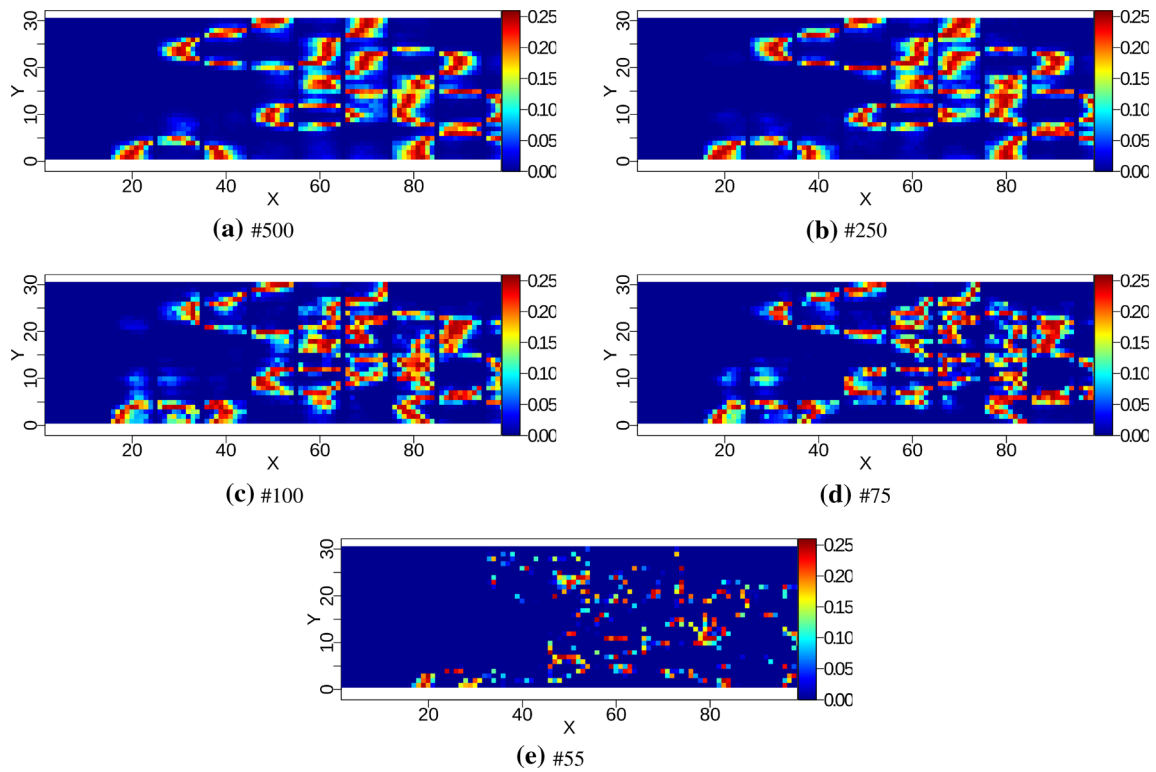


Fig. 14 Conditional variance for different number of unconditional categorical simulations corresponding to the case with less informed trend and large object size

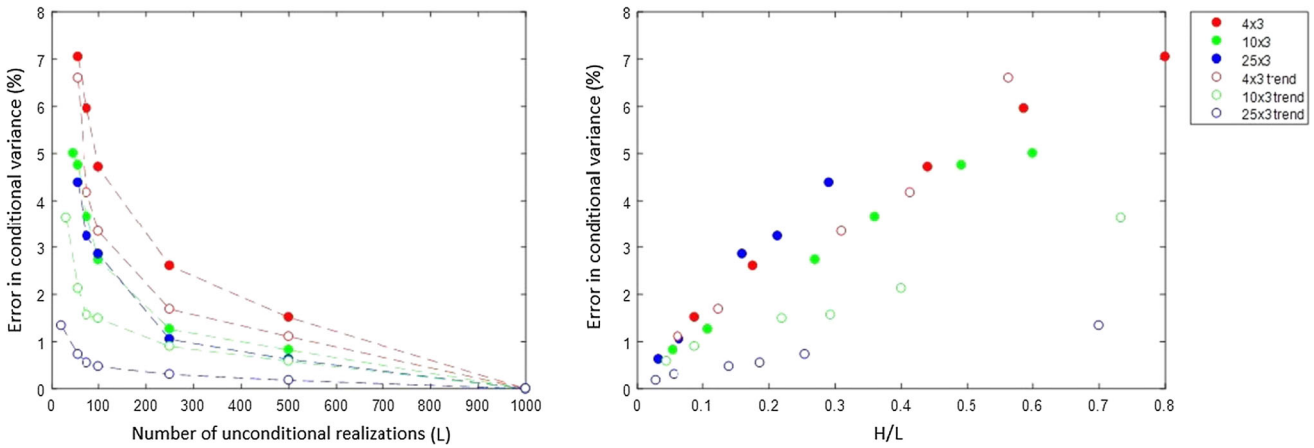


Fig. 16 Summary of the error in conditional variance for all cases (left), error in conditional variance as a function of the number of unconditional categorical realizations L and the intrinsic dimension of the data H

$$10\frac{H}{L} = error. \tag{14}$$

So to get 5% error ($error = 5$), with $H = 50$, we would need $L = 100$ unconditional simulations. It is also important to highlight that increasing the number of unconditional realizations also increases the computational time of the Gibbs sampler. In the case of more than two categories, the rule of thumb is applied to each category. Then, the relevant number of unconditional realizations is taken as the maximum of the relevant number of unconditional realizations associated with each category.

4 Real case study

In this section, the proposed conditional simulation approach is applied to real-world data. These latter are 3D lithological data from a porphyry copper deposit. A challenge in mineral resources assessment lies in estimation of boundaries between lithological, mineralization or alteration contacts. These are the $K = 6$ lithology categories, from the oldest to the youngest: wall rock (IND), Quartz Feldspar Porphyry type 1 (QFP1), Quartz Feldspar Porphyry type 2 (QFP2), intrusive breccia (ABX2), hydrothermal breccia (BRXH), and cover (COV). The physical meaning of the lithology classes is not a concern here. There are $n = 4290$ samples from 116 drill holes. The map of drill hole sample locations of interest is shown in Fig. 17. Indicator variograms and cross-variograms useful for the SIS method are shown in Fig. 18.

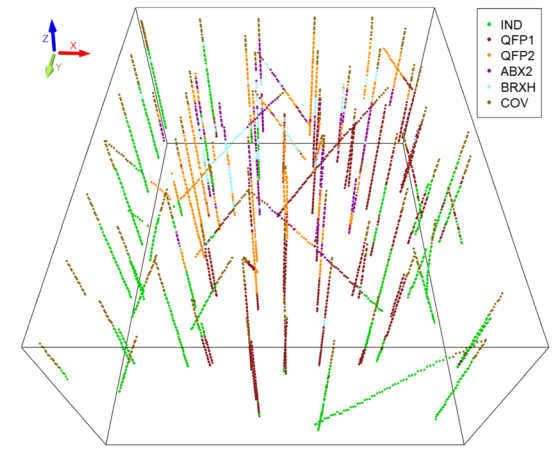


Fig. 17 Representation of lithological data

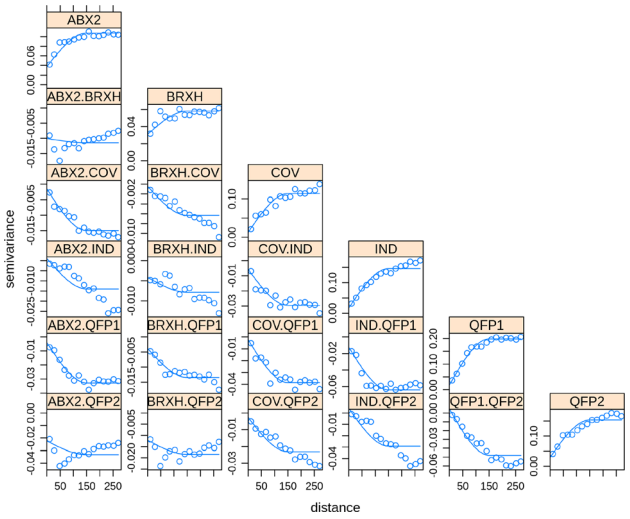


Fig. 18 Indicator variograms and cross-variograms

4.1 Unconditional categorical simulations

The proposed conditional simulation approach requires unconditional categorical simulations. Here, such

unconditional categorical simulations are generated using the following process. First, the hard data are transformed into "pseudo-signed distance" data. At data locations, the true signed distance is not observed; that would require knowing the true boundary. Instead, "pseudo signed distance" values are calculated for each data location as follows (Safa and Soltani-Mohammadi 2018):

$$\tilde{\varphi}_k(\mathbf{x}_i) = \begin{cases} -\|\mathbf{x}_i - \mathbf{x}_j\|, & \text{if } I_k(\mathbf{x}_i) = 1 \\ +\|\mathbf{x}_i - \mathbf{x}_j\|, & \text{if } I_k(\mathbf{x}_i) = 0 \end{cases}, \quad i = 1, \ldots, n; \quad k = 1, \ldots, K; \tag{15}$$

where $\mathbf{x}_j$ corresponds to the closest data location of different category than at data location $\mathbf{x}_i$. Euclidean norm is used to measure the distance. The "pseudo signed distance values" at data locations are depicted in Fig. 19.

Secondly, radial basis functions (RBF) interpolation is performed on the "pseudo signed distance" values

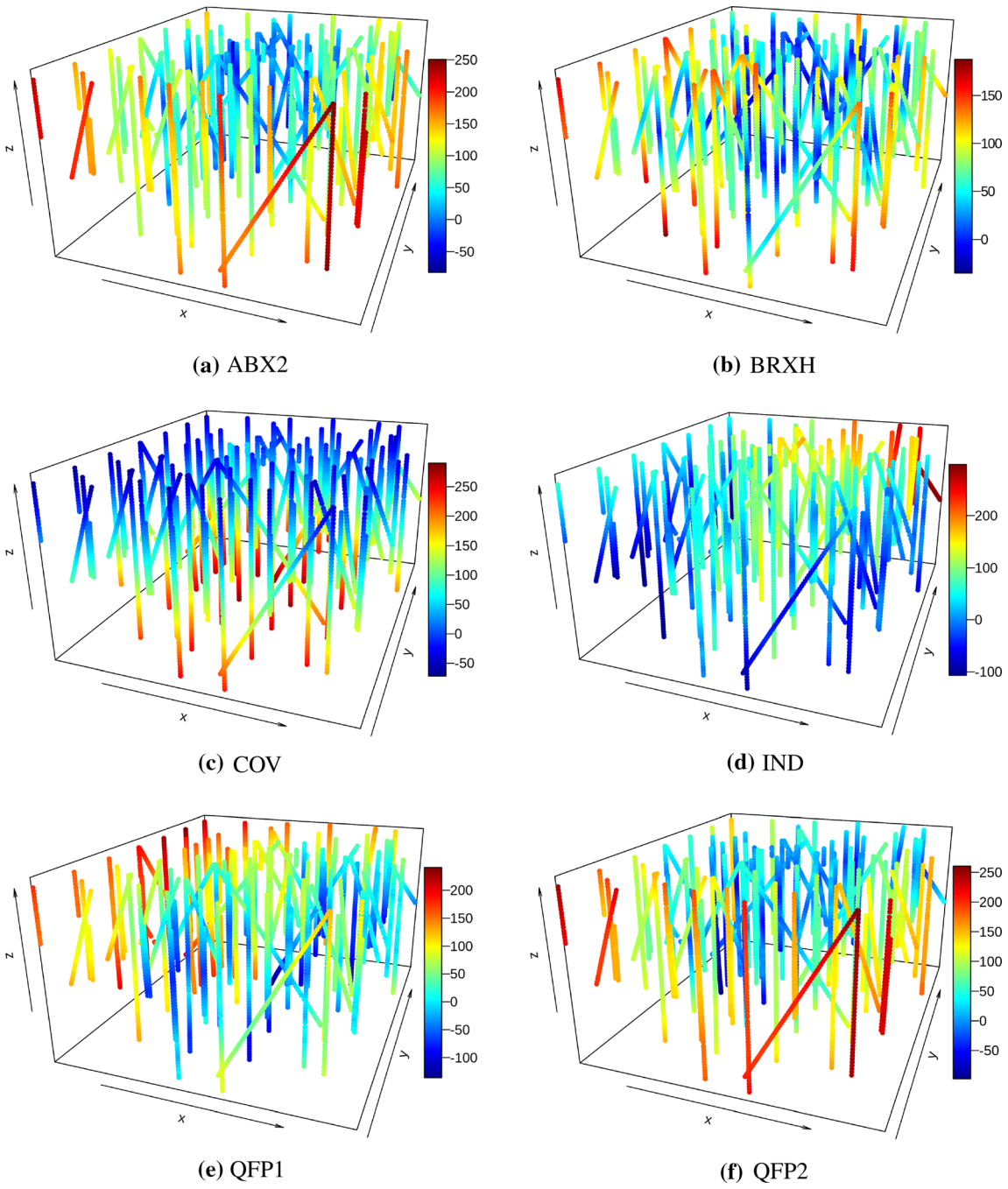


Fig. 19 Pseudo signed distance values at data locations for each lithology category

$\{\tilde{\varphi}_k(\mathbf{x}_i)\}_{i=1,\ldots,n}$ associated with each category. Specifically, one gets (Buhmann 2003):

$$\hat{\varphi}_k(\mathbf{x}) = \sum_{i=1}^{n} \omega_k^i h_k(\|\mathbf{x}-\mathbf{x}_i\|), \quad \forall \mathbf{x} \in D,\ k = 1,\ldots,K; \tag{16}$$

where the basis function $h_k(\cdot)$ is a radially-symmetric function (e.g., linear, cubic). The unknown coefficients $\boldsymbol{\omega}_k = (\omega_k^1,\ldots,\omega_k^n)$ are determined by the requirement that $\hat{\varphi}_k(\cdot)$ is an exact interpolator, i.e. $\hat{\varphi}_k(\mathbf{x}_i) = \tilde{\varphi}_k(\mathbf{x}_i)$, $i = 1,\ldots,n$. The interpolation is performed individually for each category.

Thirdly, an unconditional categorical simulation is obtaining by perturbing the RBF interpolant $\{\hat{\varphi}_k(\mathbf{x}), \mathbf{x} \in D\}$ [Eq. (16)] as follows:

$$\phi_k(\mathbf{x}) = \hat{\varphi}_k(\mathbf{x}) + \epsilon_k(\mathbf{x}),\ \forall \mathbf{x} \in D,\ k = 1,\ldots,K; \tag{17}$$

and taking $C(\mathbf{x}) = \arg\min_{\text{Gaussian}} (\varphi_1(\mathbf{x}),\ldots,\varphi_K(\mathbf{x})),\ \forall \mathbf{x} \in D$. $\epsilon_k(\cdot)$ is a zero-mean Gaussian random function with an exponential covariance function $C(\cdot;\sigma_k^2;\tau_k)$ whose parameters $\sigma_k^2 \sim \mathcal{U}(a_{1,k}, a_{2,k})$ and $\tau_k \sim \mathcal{U}(b_{1,k}, b_{2,k})$ are uncertain and uniformly distributed (sill and range, see Table 2).

Initially 1000 unconditional categorical realizations are generated. Then, the relevant number of unconditional categorical realizations to use is based on the rule of thumb given in Sect. 3.2. For each lithology category, the dimensionality of data, i.e., the number of principal components explaining 95% of the variance is illustrated in Fig. 20. Thus, the relevant number of unconditional realizations to consider is $L = 458$, following the rule of thumb. An example of four unconditional categorical simulations is given in Fig. 21. Figure 22a–c present the actual dataset $\mathbf{d}^{(0)}$ and the simulated datasets $\{\mathbf{d}^{(l)}\}_{l=1,\ldots,458}$ in the MDS space. Figure 5d shows the calculated RMD for the actual dataset $RMD(\mathbf{d}^{(0)})$ and for the simulated datasets $\{RMD(\mathbf{d}^{(l)})\}_{l=1,\ldots,458}$. As one can notice the computed RMD for the actual dataset falls below the 97.5 percentile threshold which is equal to 9.75. Thus, unconditional categorical realizations are consistent with data (Fig. 22).

Table 2 Distribution of sill and range parameters of $\epsilon_k(\cdot)$ defined in Eq. (17)

Lithology	Parameters	Distribution
ABX2	(σ_1^2, τ_1)	$\mathcal{U}(3000;4000) \times \mathcal{U}(100;135)$
BRXH	(σ_2^2, τ_2)	$\mathcal{U}(1000;2000) \times \mathcal{U}(100;135)$
COV	(σ_3^2, τ_3)	$\mathcal{U}(5000;6000) \times \mathcal{U}(100;135)$
IND	(σ_4^2, τ_4)	$\mathcal{U}(5000;6000) \times \mathcal{U}(100;135)$
QFP1	(σ_5^2, τ_5)	$\mathcal{U}(4000;5000) \times \mathcal{U}(100;135)$
QFP2	(σ_6^2, τ_6)	$\mathcal{U}(4000;5000) \times \mathcal{U}(100;135)$

4.2 Conditional categorical simulations

$T = 110{,}000$ Gibbs samples $\{\boldsymbol{\alpha}_k^{(t)}\}_{t=1,\ldots,T}$ have been drawn for each lithology category ($k = 1,\ldots,6$). The first 10,000 samples of the chain are dropped as burn-in samples, and every 100th sample is accepted. So, there are 1000 conditional simulations. The proposed method took approximately 4 hours on a desktop computer (LINUX environment) with Intel(R) Core(TM) i9-7900X CPU @ 3.30GHz (10 cores / 20 threads), and 120 GB RAM. PC scores before the conditioning (unconditional PC scores $\{\boldsymbol{\alpha}_k^{(l)}\}_{l=1,\ldots,L}$) and after the conditioning (conditional PC scores $\{\boldsymbol{\alpha}_k^{(t)}\}_{t=1,\ldots,T}$) are presented in Fig. 23. An example of 3 out of 1000 conditional categorical simulations based on the proposed approach and the SIS method is given in Fig. 24. Conditional categorical realizations providing by the proposed conditioning approach depict more regular and continuous contours than SIS conditional categorical realizations. These later show noisy features and contain artifacts that are geologically unrealistic; a well-known characteristic of the SIS method (Deutsch 1998). For instance, lithology category "QFP2" is not expected to appear above the lithology category "COV" (cover). The average proportion over 1000 conditional realizations for each lithology category is given in Table 3. Under the proposed conditional simulation method, the proportions estimated from conditional categorical realizations are close to ones estimated from the data. Whereas, under the SIS method, minor lithology categories tend to be overestimated like "ABX2" and "BRXH", corresponding to the underestimation of one of major lithology categories such as "IND" and "QFP1".

Figures 25 and 26 show respectively, the resulting conditional mean and variance maps for each lithology category under the proposed conditional simulation method and the SIS method. The general appearance of the maps of conditional mean and variance are different although showing some similar patterns. The SIS method has more uncertainty than the proposed approach due to "noisy" realization that SIS is known for (Deutsch 1998).

The proposed conditional simulation approach has been also performed for a double number of relevant unconditional categorical simulations, i.e., $L = 916$. Figure 27 show the resulting conditional mean and variance maps for each lithology category computed from 1000 conditional categorical realizations generated using this time $L = 916$ unconditional categorical realizations. The general appearance of the maps of conditional mean and variance

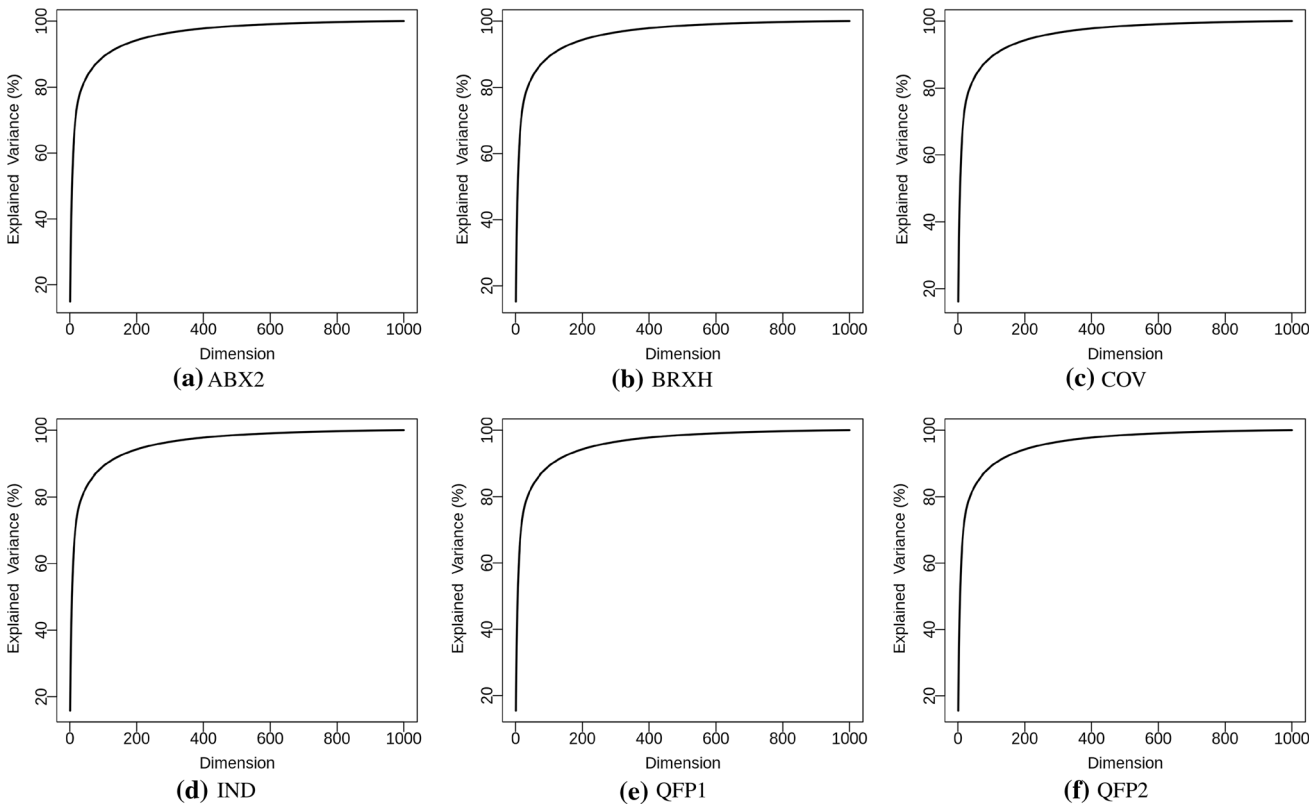


Fig. 20 Scree plot of PCA on 1000 unconditional signed distance realizations at $n = 4290$ data locations, for each lithology category

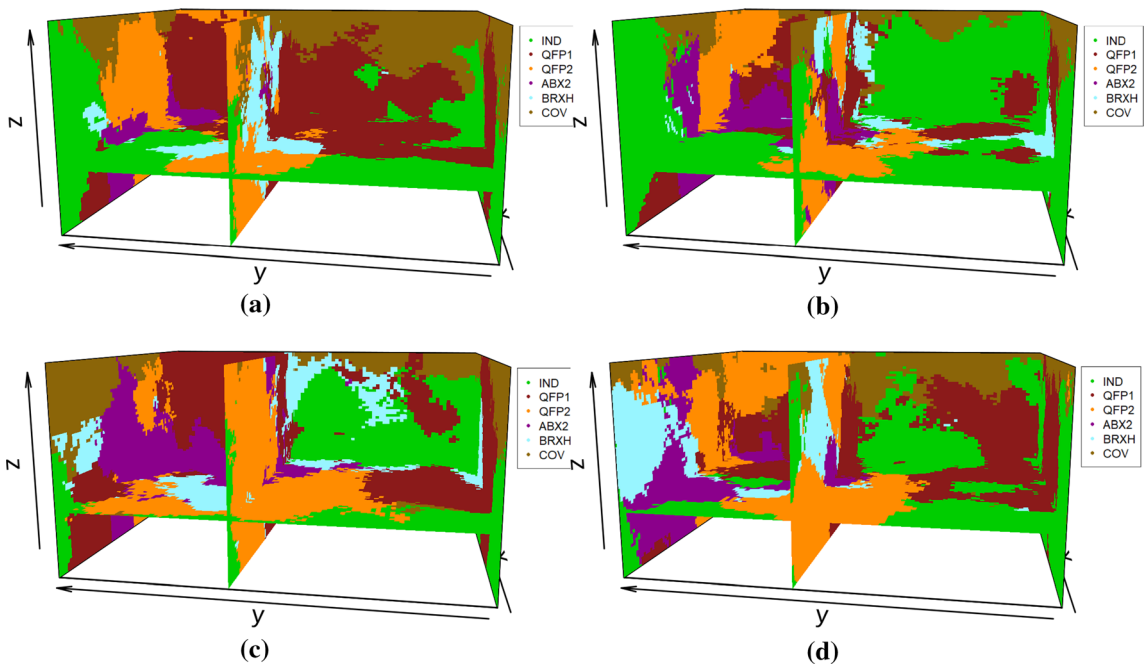
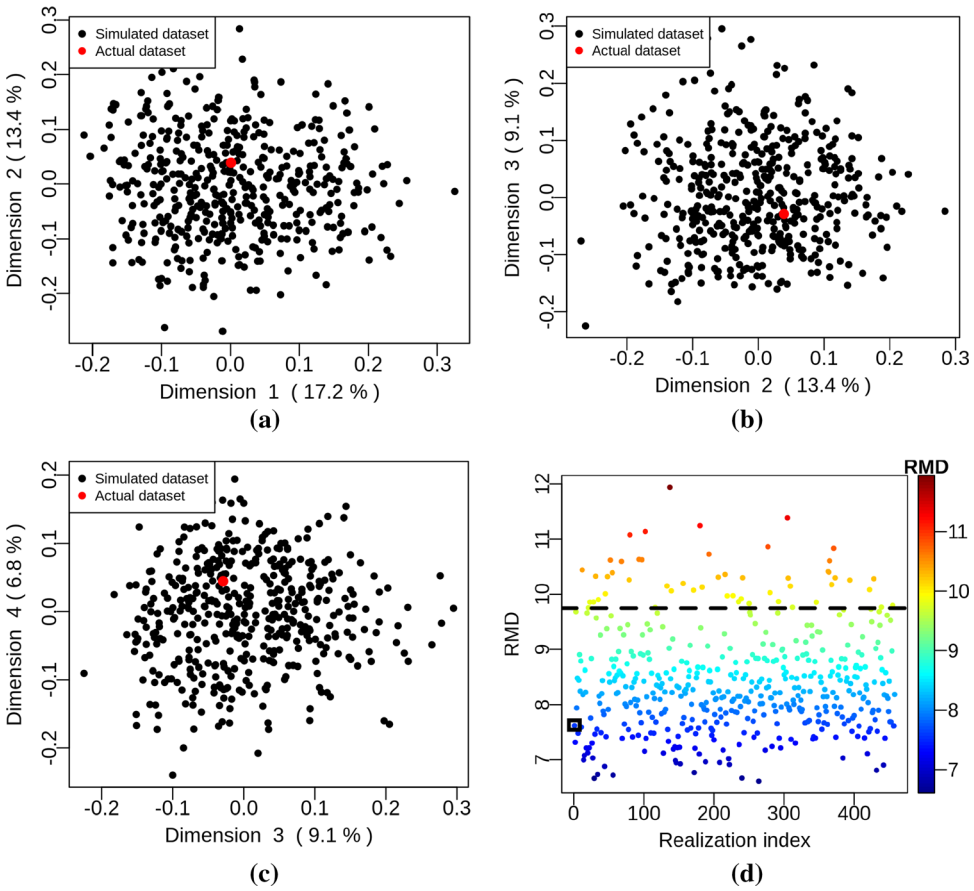


Fig. 21 Cross-sections of 4 out $L = 458$ unconditional categorical simulations

based on $L = 916$ unconditional categorical simulations is slightly different to the ones based $L = 458$ unconditional categorical simulations. This observation suggests that there is no need to use more than $L = 458$ unconditional categorical realizations to reproduce a realistic spatial uncertainty.



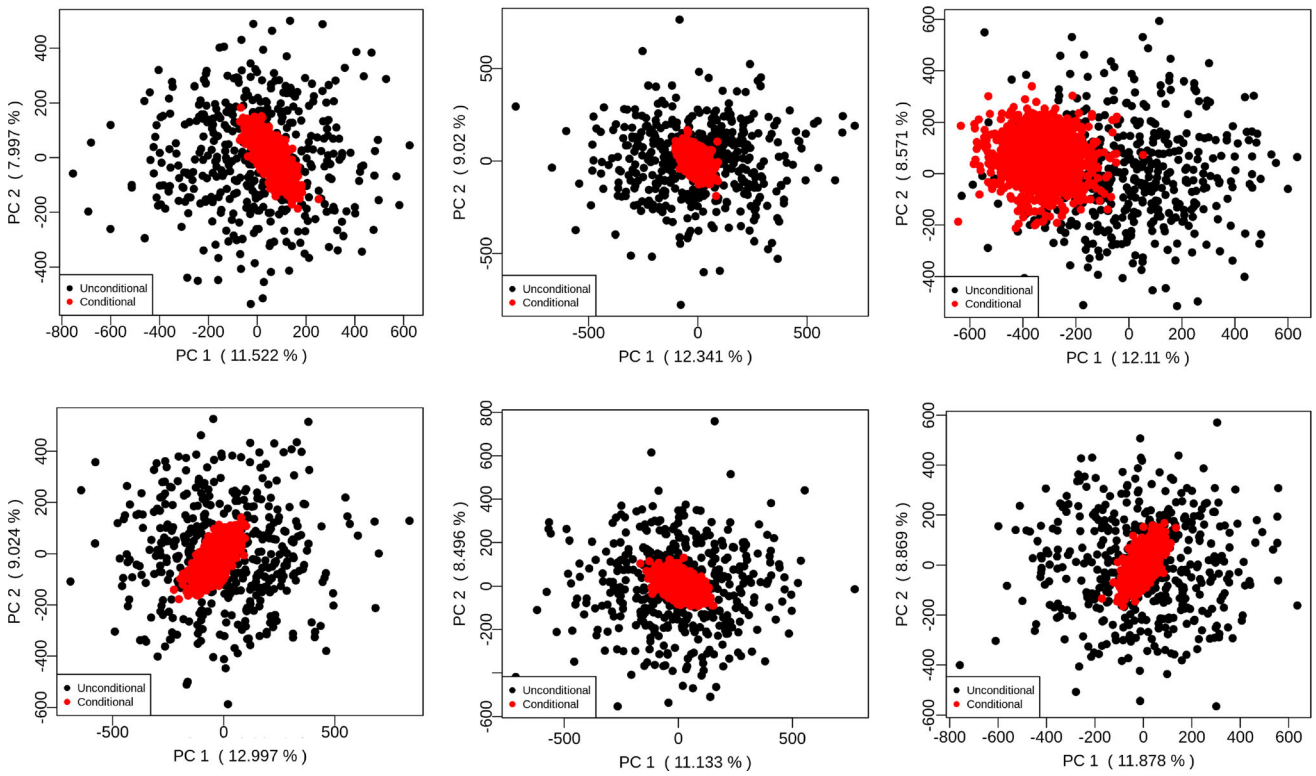


Fig. 22 Falsification of unconditional categorical simulations using robust Mahalanobis distance (RMD). **a**–**c** coordinates of datasets (actual and simulated) in the MDS space. **d** circle dots represent the calculated RMD for datasets (actual and simulated). The black-squared dot is the RMD for the actual dataset. The black dash line is the 97.5 percentile of the Chi-Squared distributed RMD

Fig. 23 $L = 458$ unconditional first PC scores and 500 conditional first PC scores for each lithology category

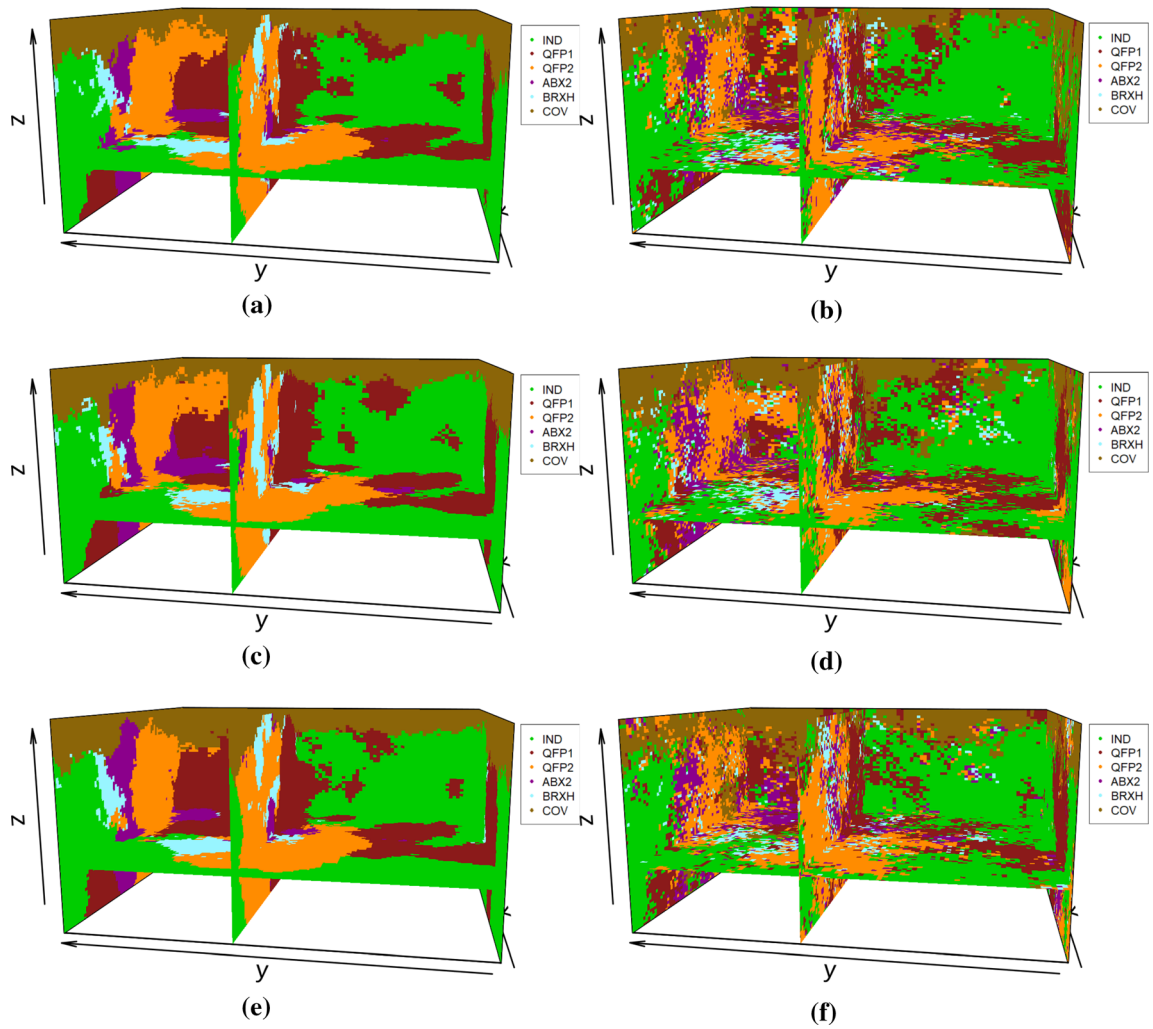


Fig. 24 **a–c** Cross-sections of 3 out 1000 conditional categorical simulations from the proposed conditioning method. **d–f** Cross-sections of 3 out 1000 conditional categorical simulations from SIS method

Table 3 Average lithology proportions over 1000 conditional simulations

	ABX2 (%)	BRXH (%)	COV (%)	IND (%)	QFP1 (%)	QFP2 (%)
Proposed method	9.46	4.76	14.59	28.42	25.41	17.36
SIS method	10.81	5.33	14.53	26.15	24.86	18.32
Data	9.64	4.59	14.65	28.16	25.76	17.20

5 Conclusions and future work

In this article, a conditioning method has been proposed for generating conditional categorical simulations from an ensemble of unconditional categorical simulations coming from any simulation approach. The proposed method takes advantage of the implicit functions representation, in combination with principal component analysis and Gibbs sampler to achieve the conditioning to the data. A rule of thumb has been derived in order to select the relevant number of unconditional categorical simulations necessary to capture the spatial uncertainty. A falsification procedure has been proposed to test the consistency between unconditional categorical simulations and the data. Synthetic and real-world case studies have been used to demonstrate the effectiveness of the proposed conditioning method.

Typical characteristics of the proposed conditional simulation approach are the following. It is independent to the method used to construct unconditional categorical simulations; it does not assume that unconditional categorical simulations are independent. The proposed method can easily handle a large number of categories in a

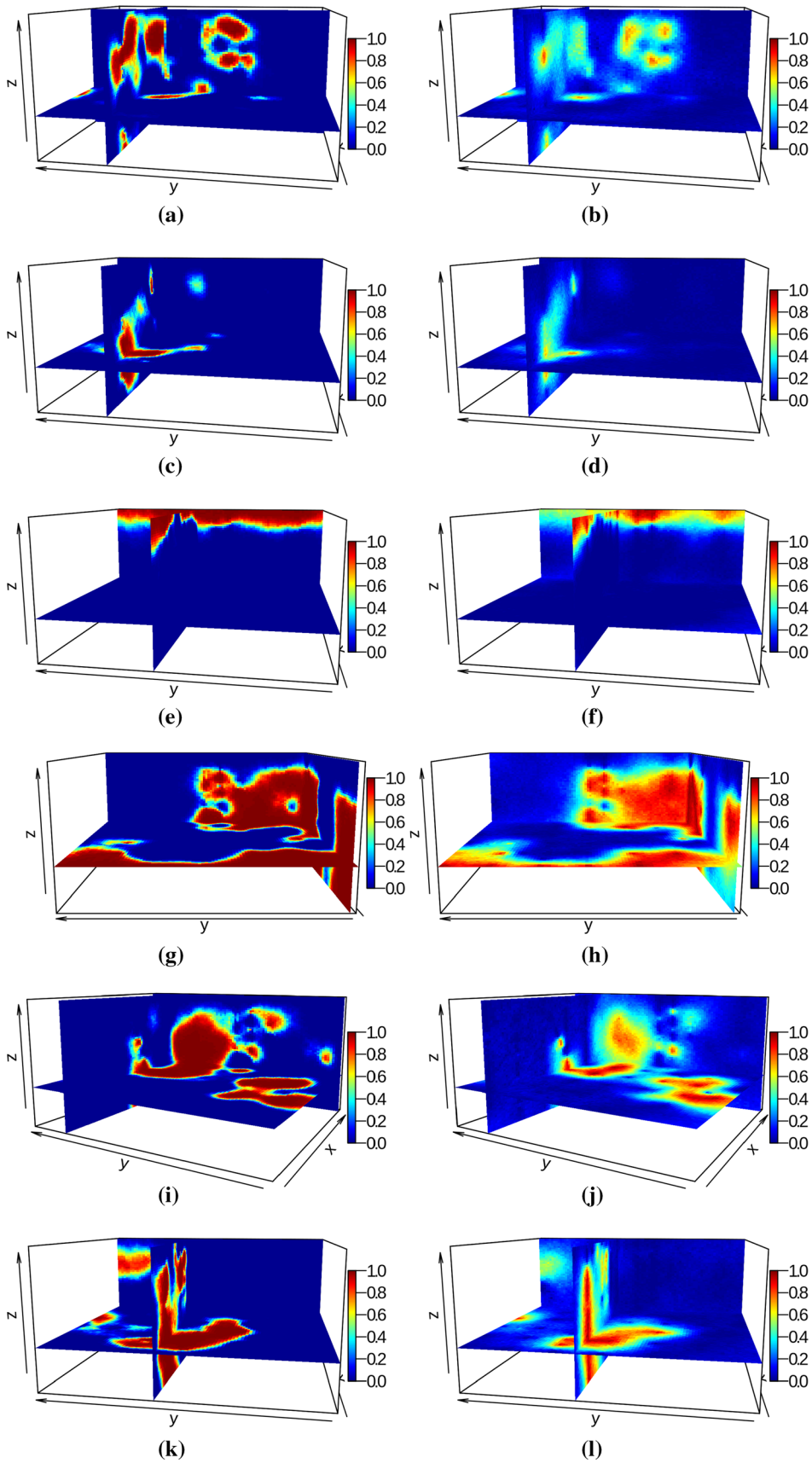


Fig. 25 Proposed method: **a**, **c**, **e**, **g**, **i**,**k** cross-sections of conditional mean for each lithology category computed from 1000 conditional simulations generated using $L = 458$ unconditional simulations. SIS method: **b**, **d**, **f**, **h**, **j**, **l** cross-sections of conditional mean for each lithology category computed from 1000 conditional simulations

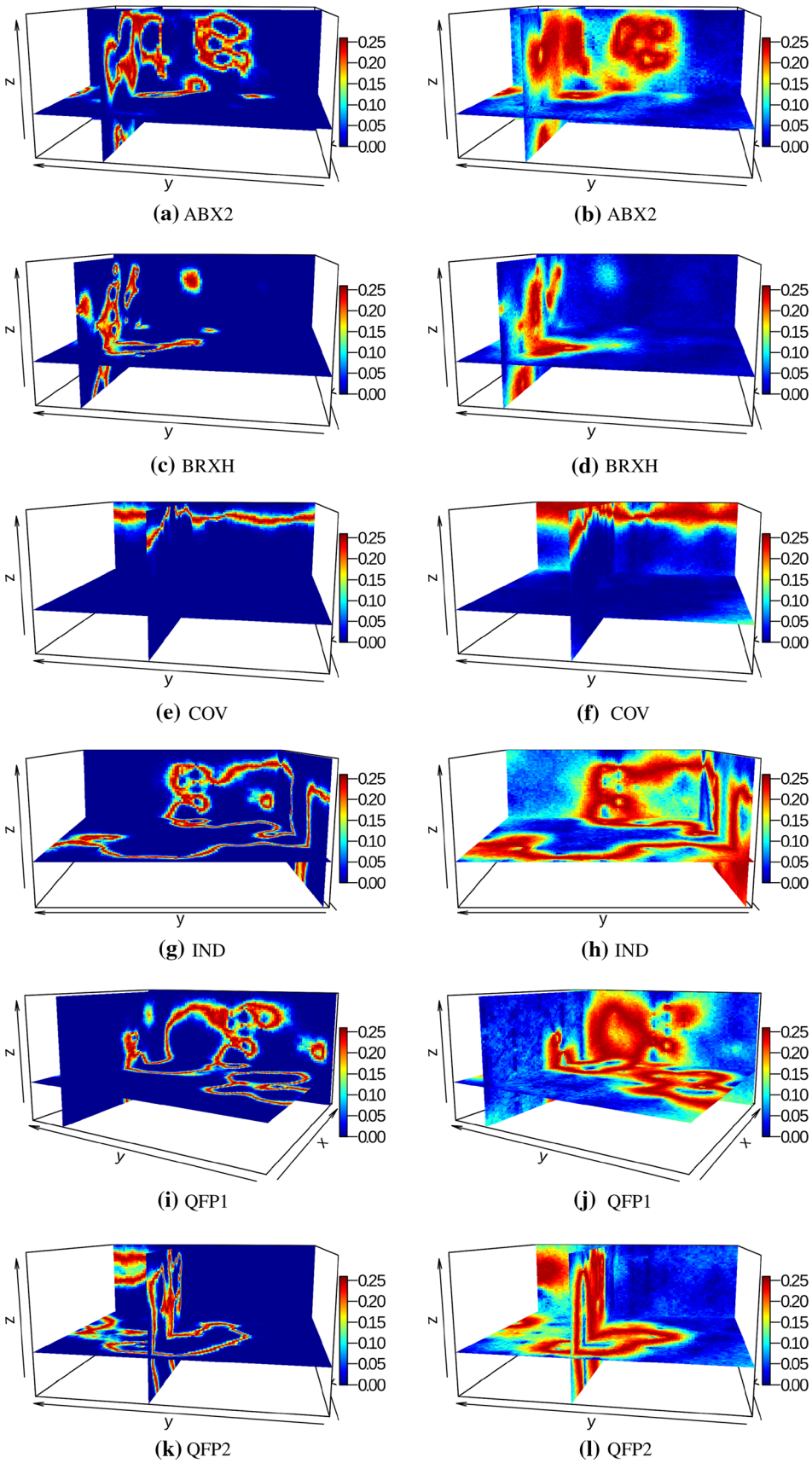


Fig. 26 Proposed method: **a**, **c**, **e**, **g**, **i**, **k** cross-sections of conditional variance for each lithology category computed from 1000 conditional simulations generated using $L = 458$ unconditional simulations. SIS method: **b**, **d**, **f**, **h**, **j**, **l** cross-sections of conditional variance for each lithology category computed from 1000 conditional simulations

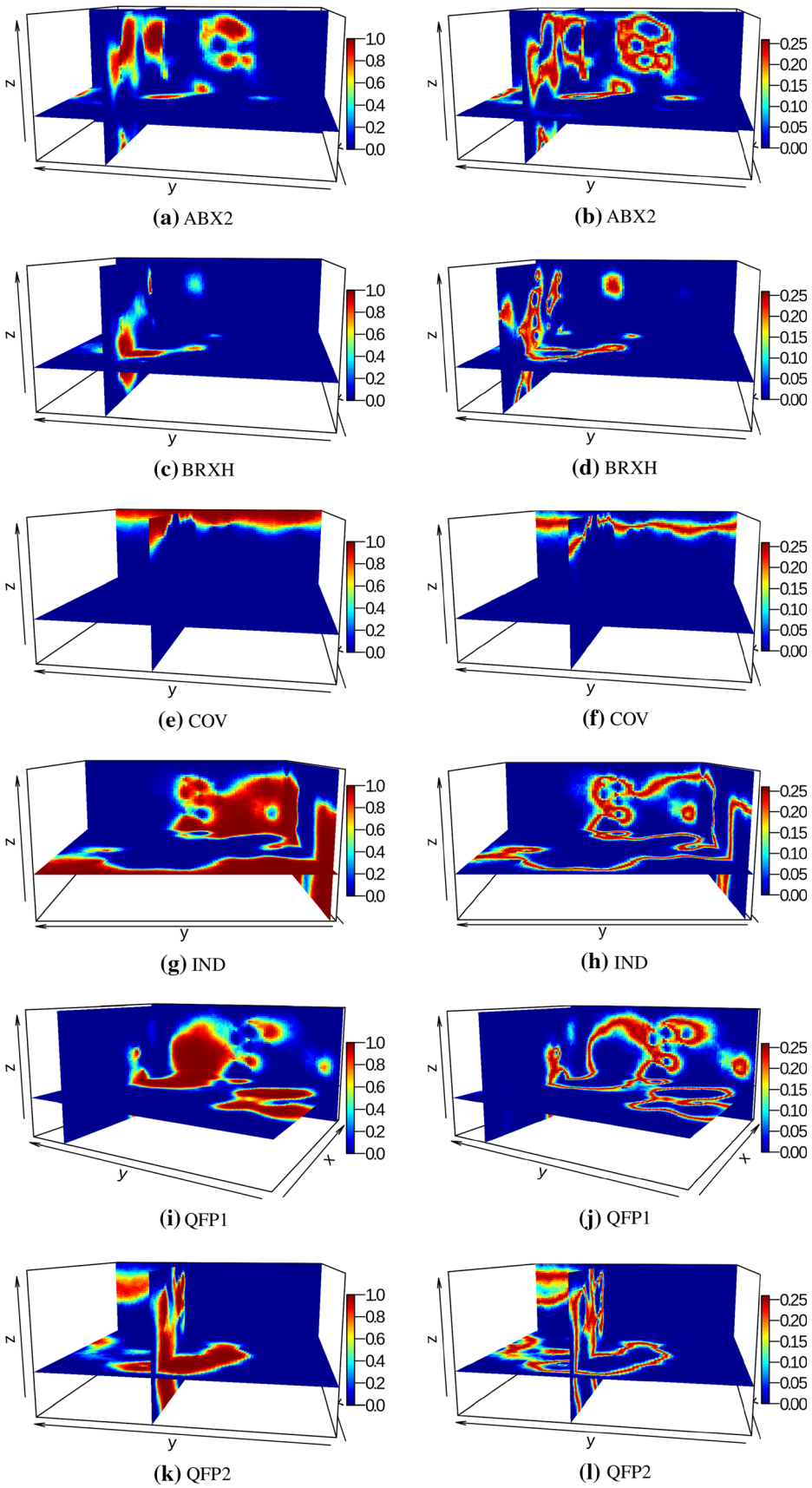


Fig. 27 **a**, **c**, **e**, **g**, **i**, **k** cross-sections of conditional mean for each lithology category computed from 1000 conditional simulations generated using $L = 916$ unconditional simulations. **b**, **d**, **f**, **h**, **j**, **l** conditional variance for each lithology category computed from 1000 conditional simulations generated using $L = 916$ unconditional simulations

consistent manner via the implicit function representation; it can be applied when categories obey an ordered sequence like stratigraphies and lithologies; this order relation can be captured through signed distance functions that are truncated according to a set of rules. The proposed technique can be performed in any dimension (e.g., 1D, 2D and 3D); it can be carried out when the conditioning data are very irregularly or regularly located. The proposed approach provides more realistic categorical realizations than the SIS method as shown in the real case study. It comprises some components that can be performed in parallel according to the number of categories, including the generation of conditional PC scores.

The proposed conditional simulation method relies on the Gibbs sampling of a truncated multivariate Gaussian distribution subject to linear inequality constraints. For each category, the number of linear inequality constraints is equal to the number of data points. The computational time of the proposed method increases with the grid size, the number of unconditional categorical simulations, and the number the data points. When dealing with very large datasets, the proposed method could be time consuming as many conditional simulation methods. To overcome this problem, the number of linear constraints can be reduced due to some redundancy existing in very large datasets. Specifically, very large datasets often exhibit clustered data points. For clustered data points with the same category, only few data points can be considered to derive linear inequality constraints without affecting the conditioning.

Under the proposed conditional simulation technique, the PC scores are assumed to follow the normal distribution. So, prior to apply the proposed approach, the normality assumption of the unconditional PC scores should be checked. In case where the unconditional PC scores deviated significantly from the normal distribution, the resulting conditional categorical simulations might not reproduce some statistical properties. It would be interesting to extend the proposed method to accommodate other distributions other than the Gaussian distribution.

Acknowledgements We gratefully acknowledge the funding provided by BHP to support this research. We are grateful to the anonymous reviewers for their helpful and constructive comments that greatly helped improve the manuscript.

References

Armstrong M, Galli A, Beucher H, Loc'h G, Renard D, Doligez B, Eschard R, Geffroy F (2011) Plurigaussian simulations in geosciences. Springer, Berlin

Arpat GB, Caers J (2007) Conditional simulation with patterns. Math Geol 39(2):177–203

Bogaert P, Gengler S (2018) Bayesian maximum entropy and data fusion for processing qualitative data: theory and application for crowdsourced cropland occurrences in ethiopia. Stoch Env Res Risk Assess 32(3):815–831

Borg I, Groenen P (2007) Modern multidimensional scaling: theory and applications. Springer Series in Statistics. Springer, New York

Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) Lof: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data, pp 93–104

Buhmann M (2003) Radial basis functions: theory and implementations. Cambridge monographs on applied and computational mathematics. Cambridge University Press, Cambridge

Chiles J-P, Delfiner P (2012) Geostatistics: modeling spatial uncertainty. Wiley, New York

Daly C (2005) Higher Order models using entropy. Markov random fields and sequential simulation. Springer, Dordrecht, pp 215–224

Davies E (2012) Chapter 9—binary shape analysis. In: Davies E (ed) Computer and machine vision, 4th edn. Academic Press, Boston, pp 229–265

Deutsch C (2002) Geostatistical reservoir modelling. Oxford University Press, Oxford

Deutsch CV (1998) Cleaning categorical variable (lithofacies) realizations with maximum a-posteriori selection. Comput Geosci 24(6):551–562

Deutsch CV (2006) A sequential indicator simulation program for categorical variables with point and block data: blocksis. Comput Geosci 32(10):1669–1681

Elfeki A, Dekking M (2001) A markov chain model for subsurface characterization: theory and applications. Math Geol 33(5):569–589

Emery X (2007) Simulation of geological domains using the plurigaussian model: new developments and computer programs. Comput Geosci 33(9):1189–1201

Goovaerts P (1998) Geostatistics for natural resources evaluation. Oxford University Press, Oxford

Grevera GJ (2007) Distance transform algorithms and their implementation and evaluation. Springer, New York, pp 33–60

Honarkhah M, Caers J (2010) Stochastic simulation of patterns using distance-based pattern modeling. Math Geosci 42(5):487–517

Horrace WC (2005) Some results on the multivariate truncated normal distribution. J Multivar Anal 94(1):209–221

Hubert M, Debruyne M, Rousseeuw PJ (2018) Minimum covariance determinant and extensions. WIREs Comput Stat 10(3):e1421

Journel AG (1983) Nonparametric estimation of spatial distributions. J Int Assoc Math Geol 15(3):445–468

Lantuejoul C (2002) Geostatistical simulation: models and algorithms. Springer, New York

Li W (2007) Markov chain random fields for estimation of categorical variables. Math Geol 39(3):321–335

Li Y, Ghosh SK (2015) Efficient sampling methods for truncated multivariate normal and student-t distributions subject to linear inequality constraints. J Stat Theory Pract 9(4):712–732

Liu FT, Ting KM, Zhou Z (2008) Isolation forest. In: 2008 Eighth IEEE international conference on data mining, pp 413–422

Mariethoz G, Caers J (2014) Multiple-point geostatistics: stochastic modeling with training images. Willey, New York

Mariethoz G, Renard P, Straubhaar J (2010) The direct sampling method to perform multiple-point geostatistical simulations. Water Resour Res 46(11):1–14

Osher S, Fedkiw R (2002) Level set methods and dynamic implicit surfaces. Applied mathematical sciences. Springer, New York

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed 17 July 2020

Safa M, Soltani-Mohammadi S (2018) Distance function modeling in optimally locating additional boreholes. Spatial Stat 23:17–35
Scheidt C, Li L, Caers J (2018) Quantifying uncertainty in subsurface systems. Geophysical monograph series. Wiley, New York
Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. Neural Comput 13(7):1443–1471
Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. Math Geol 34(1):1–21
Tjelmeland H, Besag J (1998) Markov random fields with higher-order interactions. Scand J Stat 25(3):415–433
Vanderbei R (2013) Linear programming: foundations and extensions. International series in operations research & management science. Springer, New York
Yang L, Hyde D, Grujic O, Scheidt C, Caers J (2019) Assessing and visualizing uncertainty of 3D geological surfaces using level sets with stochastic motion. Comput Geosci 122:54–67
Zhang T, Switzer P, Journel A (2006) Filter-based classification of training image patterns for spatial simulation. Math Geol 38(1):63–80
Zhou Y, Zhang W, Zhu J, Xu Z (2016) Feature-driven topology optimization method with signed distance function. Comput Methods Appl Mech Eng 310:1–32
Žukovič M, Hristopulos DT (2009) Classification of missing values in spatial data using spin models. Phys Rev E 80:011116