

A new computational model of high-order stochastic simulation based on spatial Legendre moments

L. Yao. R. Dimitrakopoulos,
M. Gamache

G-2018-89

October 2018

La collection *Les Cahiers du GERAD* est constituée des travaux de recherche menés par nos membres. La plupart de ces documents de travail a été soumis à des revues avec comité de révision. Lorsqu'un document est accepté et publié, le pdf original est retiré si c'est nécessaire et un lien vers l'article publié est ajouté.

Citation suggérée: L. Yao. R. Dimitrakopoulos, M. Gamache (Octobre 2018). A new computational model of high-order stochastic simulation based on spatial Legendre moments, Rapport technique, Les Cahiers du GERAD G-2018-89, GERAD, HEC Montréal, Canada.

Avant de citer ce rapport technique, veuillez visiter notre site Web (<https://www.gerad.ca/fr/papers/G-2018-89>) afin de mettre à jour vos données de référence, s'il a été publié dans une revue scientifique.

The series *Les Cahiers du GERAD* consists of working papers carried out by our members. Most of these pre-prints have been submitted to peer-reviewed journals. When accepted and published, if necessary, the original pdf is removed and a link to the published article is added.

Suggested citation: L. Yao. R. Dimitrakopoulos, M. Gamache (October 2018). A new computational model of high-order stochastic simulation based on spatial Legendre moments, Technical report, Les Cahiers du GERAD G-2018-89, GERAD, HEC Montréal, Canada.

Before citing this technical report, please visit our website (<https://www.gerad.ca/en/papers/G-2018-89>) to update your reference data, if it has been published in a scientific journal.

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2018
– Bibliothèque et Archives Canada, 2018

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2018
– Library and Archives Canada, 2018

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Télec. : 514 340-5665
info@gerad.ca
www.gerad.ca

A new computational model of high-order stochastic simulation based on spatial Legendre moments

Lingqing Yao ^{a,b,c}

Roussos Dimitrakopoulos ^{a,b}

Michel Gamache ^{a,c}

^a GERAD, Montréal (Québec), Canada, H3T 2A7

^b COSMO – Stochastic Mine Planning Laboratory
& Department of Mining and Materials Engineering,
McGill University, Montréal (Québec) Canada,
H3A 0E8

^c Department of Mathematics and Industrial Engineering,
Polytechnique Montréal (Québec) Canada,
H3C 3A7

lingqing.yao@mcgill.ca

roussos.dimitrakopoulos@mcgill.ca

michel.gamache@polymtl.ca

October 2018

Les Cahiers du GERAD

G–2018–89

Copyright © 2018 GERAD, Yao. Dimitrakopoulos, Gamache

Les textes publiés dans la série des rapports de recherche *Les Cahiers du GERAD* n'engagent que la responsabilité de leurs auteurs. Les auteurs conservent leur droit d'auteur et leurs droits moraux sur leurs publications et les utilisateurs s'engagent à reconnaître et respecter les exigences légales associées à ces droits. Ainsi, les utilisateurs:

- Peuvent télécharger et imprimer une copie de toute publication du portail public aux fins d'étude ou de recherche privée;
- Ne peuvent pas distribuer le matériel ou l'utiliser pour une activité à but lucratif ou pour un gain commercial;
- Peuvent distribuer gratuitement l'URL identifiant la publication.

Si vous pensez que ce document enfreint le droit d'auteur, contactez-nous en fournissant des détails. Nous supprimerons immédiatement l'accès au travail et enquêterons sur votre demande.

The authors are exclusively responsible for the content of their research papers published in the series *Les Cahiers du GERAD*. Copyright and moral rights for the publications are retained by the authors and the users must commit themselves to recognize and abide the legal requirements associated with these rights. Thus, users:

- May download and print one copy of any publication from the public portal for the purpose of private study or research;
- May not further distribute the material or use it for any profit-making activity or commercial gain;
- May freely distribute the URL identifying the publication.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Abstract: Multiple-point simulations have been introduced over the past decade to overcome the limitations of second-order stochastic simulations in dealing with geologic complexity, curvilinear patterns and non-Gaussianity. However, a limitation is that they sometimes fail to generate results that comply with the statistics of the available data while maintaining the consistency of high-order spatial statistics. As an alternative, high-order stochastic simulations based on spatial cumulants or spatial moments have been proposed; however, they are also computationally demanding, which limits their applicability.

The present work derives a new computational model to numerically approximate the conditional probability density function (*cpdf*) as a multivariate Legendre polynomial series based on the concept of spatial Legendre moments. The advantage of this method is that no explicit computations of moments (or cumulants) are needed in the model. The approximation of the *cpdf* is simplified to the computation of a unified empirical function. Moreover, the new computational model computes the *cpdfs* within a local neighborhood without storing the high-order spatial statistics through a predefined template. With this computational model, the algorithm for the estimation of the *cpdf* is developed in such a way that the conditional cumulative distribution function (*ccdf*) can be computed conveniently through another recursive algorithm. In addition to the significant reduction of computational cost, the new algorithm maintains higher numerical precision compared to the original version of the high-order simulation. A new method is also proposed to deal with the replicates in the simulation algorithm, reducing the impacts of conflicting statistics between the sample data and the training image. A brief description of implementation is provided, and for comparison and verification, a set of case studies are conducted and compared with the results of the well-established multi-point simulation algorithm, *filtersim*. This comparison demonstrates that the proposed high-order simulation algorithm can generate spatially complex geological patterns while also reproducing the high-order spatial statistics from the sample data.

Keywords: High-order stochastic simulation, multi-point statistics, spatial moments, Legendre polynomials

1 Introduction

For the past several decades, stochastic simulations have been used to quantify spatial uncertainty in earth science applications. Traditionally, stochastic models are built on the basis of the Gaussian distribution and two-point statistics, where covariance or variograms are used to capture the spatial correlations (David 1988; Deutsch and Journel 1992; Journel 1994; Goovaerts 1997). The limitations of the existing two-point simulation methods have been reported in various publications (Guardiano and Srivastava 1993; Xu 1996; Journel 1997; Journel 2003; De Iaco and Maggio 2011), which are mostly related to the poor reproduction of spatial distributions while dealing with the complex spatial patterns, spatial connectivity of extreme values and non-Gaussianity. To reflect the complex geological patterns, multi-point statistics have to be introduced instead of conventional two-point statistics. Guardiano and Srivastava (1993) propose a multiple-point simulation (*mps*) framework and the concept of the training image (TI). The primary difference between *mps* and two-point simulations is that the conditional cumulative distribution functions (*ccdf*) are built on empirical estimations of conditional probabilities with multiple-point configurations, which is equivalent to solving a normal equation according to the Bayes' rule. Strebelle (2002) formalizes the method and develops the first computationally-efficient implementation. For over a decade, research has been focused on various issues around *mps* algorithms, such as computational efficiency and various patch-based extensions (Zhang et al. 2006; Arpat and Caers 2007; Wu et al. 2008; Boucher 2009; Remy et al. 2009; Honarkhah and Caers 2010; Mariethoz et al. 2010; Parra and Ortiz 2011; Huang et al. 2013; Boucher et al. 2014; Strebelle and Cavelius 2014; Chatterjee et al. 2016; Li et al. 2016). In general, these *mps* methods are TI-based, and their statistics are estimated from distributions of replicates of data events in the training image. Their main drawbacks are: (1) the high-order statistics are partially and indirectly considered; (2) the methods are not driven by a consistent mathematical framework; and, (3) since they are TI-driven, they may not generate results that comply with the statistics of actual available data. The latter shortcoming becomes distinctly clear in mining applications, where dense data sets are used (Osterholt and Dimitrakopoulos 2007; Goodfellow et al. 2012).

As an alternative, a high-order simulation framework with mathematical consistency is proposed with the introduction of a new concept of spatial cumulants (Dimitrakopoulos et al. 2010). The so-called high-order simulation algorithm (*hosim*) and its implementation are developed by Mustapha and Dimitrakopoulos (2010b, 2011). In this algorithm, the conditional probability density function (*cpdf*) is approximated by a multivariate expansion with coefficients expressed in terms of spatial cumulants. The *hosim* algorithm is extended mostly recently to deal with the joint simulation of multiple variables as well as the simulation of categorical data (Minniakhmetov and Dimitrakopoulos 2017a, b); other extensions are approximating the *cpdf* with different types of orthogonal polynomial bases, such as expansion series with Laguerre polynomials and Legendre-Like spline polynomials (Mustapha and Dimitrakopoulos 2010a; Minniakhmetov and Dimitrakopoulos 2018). However, the related calculations are computationally demanding, since the number of spatial cumulants involved in the series increases exponentially either as the order of cumulants or the quantity of conditioning data increases. In Mustapha and Dimitrakopoulos (2011), some terms of the expansion series have to be discarded to obtain computational feasibility, which compromises the accuracy of the approximated *cpdf*. In addition, the computational cost limits the approach for larger-scale applications.

To take full advantage of the high-order simulation, that is, its data-driven aspect and no presumption of data distribution, and address the computational difficulties, a new stochastic simulation algorithm based on high-order spatial Legendre moments is presented herein. Rather than just a mathematical equivalency of the previous model of the high-order simulation, the approximation of *cpdf* by Legendre polynomial series is reformulated under the framework of the sequential simulation, leading to a much more concise form of the computational model. In this new method, all explicit calculations of moments are encapsulated in a unified function to derive the *cpdf*, cutting down the previous complex computations into a few iterations of simple operations with polynomial time. Moreover, there is no predefined template configuration in the new algorithm, as required for the normal *mps* methods and the previous *hosim* model. The spatial configuration of the template will instead depend on the local neighborhood of the node to be simulated; note that there is no need to store the intermediate results in a tree as in most of the *mps* methods including the previous *hosim*. The variable template also has the advantage of simultaneously capturing the spatial patterns either in local scale or global scale.

The remainder of the paper continues with Section 2, which describes the stochastic model based on the concepts of high-order spatial Legendre moments. Section 3 develops the computational model as a statistical function. Section 4 describes the new proposed high-order simulation algorithm and analyzes the computational complexity. Section 5 explores the implementation of the new high-order simulation algorithm. Section 6 shows the examples to assess the new method and compare it with *filtersim*. Finally, conclusions and future research are presented in Section 7.

2 Stochastic model of high-order simulation with spatial Legendre moments

2.1 Sequential simulation

In this paper, the stochastic model is discussed specifically under the sequential simulation framework (Rosenblatt 1952; Johnson 1987; Journel 1994). Sequential simulation aims to reproduce spatial properties sequentially by decomposing the multivariate conditional distributions into a set of univariate distributions. Considering a stationary and ergodic random field $\mathbf{Z}(\mathbf{u})$, let $Z(\mathbf{u}_1), \dots, Z(\mathbf{u}_N)$ be a set of random variables with locations at $\mathbf{u}_1, \dots, \mathbf{u}_N$, respectively. Then, the N random variables $Z(\mathbf{u}_1), \dots, Z(\mathbf{u}_N)$ constitute a joint multivariate distribution. In terms of the stochastic simulation, it is supposed that realizations are to be generated from $Z(\mathbf{u}_1), \dots, Z(\mathbf{u}_N)$, and the available data set are $\Lambda_0 = \{\zeta(\mathbf{u}'_1), \dots, \zeta(\mathbf{u}'_n)\}$, where $\zeta(\mathbf{u}'_i)$ is the sample data at the location $\mathbf{u}'_i$ for $i = 1, \dots, n$, and n is the number of sample data in total. For simplification, $Z(\mathbf{u}_1), \dots, Z(\mathbf{u}_N)$ are alternatively written as $Z_1, \dots, Z_N$, and a similar simplification of notation applies in the context of a random field. Following the above notation, the stochastic simulation of the random field is based on the sampling from the N -variate probability distribution posterior to the data set Λ_0 , which can be characterized by a conditional cumulative distribution function (*ccdf*) as $F_{\mathbf{Z}}(z_1, \dots, z_N | \Lambda_0)$ or by a probability density function (*cpdf*) as $f_{\mathbf{Z}}(z_1, \dots, z_N | \Lambda_0)$. The joint *cpdf* $f_{\mathbf{Z}}(z_1, \dots, z_N | \Lambda_0)$ can be decomposed into the product of a series of univariate *cpdfs* (Rosenblatt 1952; Johnson 1987) as

$$f_{\mathbf{Z}}(z_1, \dots, z_N | \Lambda_0) = f_{Z_1}(z_1 | \Lambda_0) \cdots f_{Z_N}(z_N | \Lambda_{N-1}), \quad (1)$$

where $\Lambda_i (i = 1, \dots, N-1)$ are a series of sets and $\Lambda_i = \Lambda_{i-1} \cup \{\zeta(\mathbf{u}_i)\}$, $i = 1, \dots, N$, where $\zeta(\mathbf{u}_i)$ is the value drawn from the conditional probability distribution with a density function described as $f_{Z_i}(z_i | \Lambda_0)$.

The basic idea of sequential simulation is to sequentially draw random values from the decomposed univariate *cpdfs* through a random path that visits all the nodes to be simulated. Irrespective of the node's location corresponding to the sequence number, there is no difference in the sampling procedures. Without loss of generality, the *cpdf* in every single sampling procedure can be symbolized uniformly as $f_{Z_0}(z_0 | \Lambda)$, where Z_0 means the current simulating node and Λ means the set of conditioning data around Z_0 's location $\mathbf{u}_0$. Considering the computational intensity and the statistical relevancy, the conditioning data are usually confined to a neighborhood closest to the simulation node instead of taking account of all available data on the whole domain of the random field. For more details on this screen-effect approximation, the reader is referred to Dimitrakopoulos and Luo (2004).

An algorithmic description of sequential simulation can be summarized as the following steps:

1. Draw a random path to visit all the N nodes to be simulated.
2. Starting from $i = 1$ and for each node $Z(\mathbf{u}_i)$, derive the conditional probability cumulative distribution $F_{Z_i}(z_i | \Lambda_{i-1})$ or the density function $f_{Z_i}(z_i | \Lambda_{i-1})$.
3. Draw a random value $\zeta(\mathbf{u}_i)$ from the conditional probability distribution in Step 2 and update the conditioning data by adding the node value $\zeta(\mathbf{u}_i)$ in to the current data set Λ_i .
4. Repeat from Step 2 until all the nodes are visited.

2.2 High-order spatial Legendre moments

In probability theory, moments are defined as expectations of integer power functions of a random variable. Given a random variable Z in probability space (Ω, F, P) , suppose that the density of probability measure P is a continuous function $f_Z(z)$. The moment of order w is defined as

$$\text{Mom}_Z(w) = E[Z^w] = \int_{\Omega} [Z(\omega)]^w dP(\omega) = \int_{\mathbb{R}} z^w f_Z(z) dz. \quad (2)$$

The moments of random vector $\mathbf{Z} = [Z_0, \dots, Z_N]$ with a multivariate density $f_{\mathbf{Z}}(z_0, \dots, z_N)$ defined similarly as

$$\text{Mom}_{\mathbf{Z}}(w_0, \dots, w_N) = E[z_1^{w_0} \dots z_N^{w_N}] = \int_{\mathbb{R}^N} z_0^{w_0} \dots z_N^{w_N} f_{\mathbf{Z}}(z_0, \dots, z_N) dz_0 \dots dz_N, \quad (3)$$

where $w_i (i=0, \dots, N)$ are the orders of moments for the i -th element of vector $\mathbf{Z}$. The spatial moments of a discrete random field $\mathbf{Z} = [Z(\mathbf{u}_0), \dots, Z(\mathbf{u}_n)]$ are functions of spatial location variables $\mathbf{u}_0, \dots, \mathbf{u}_n$. Assuming the random field $\mathbf{Z}(\mathbf{u})$ is stationary and ergodic, the spatial moments of $\mathbf{Z}(\mathbf{u})$ can be expressed as functions of distance vectors, and thus they are independent of the locations. These distance vectors, which keep the spatial configuration of a center node and nodes within its neighborhood, can be expressed using a spatial template $\mathbf{T}$ (Figure 1). The terminologies of the spatial template $\mathbf{T}$ and data events (Strebelle 2002; Dimitrakopoulos et al. 2010) are the following:

- (i) Spatial template $\mathbf{T}$: geometry defined by N distance vectors $(\mathbf{h}_1, \dots, \mathbf{h}_N)$ from the center node $\mathbf{u}_0$, $\mathbf{T} = \{\mathbf{u}_0, \mathbf{u}_0 + \mathbf{h}_1, \dots, \mathbf{u}_0 + \mathbf{h}_N\}$.
- (ii) Data events: outcomes of the random field in the spatial template $\mathbf{T}$. Specifically, the data events are conditioning data set Λ in the present work.

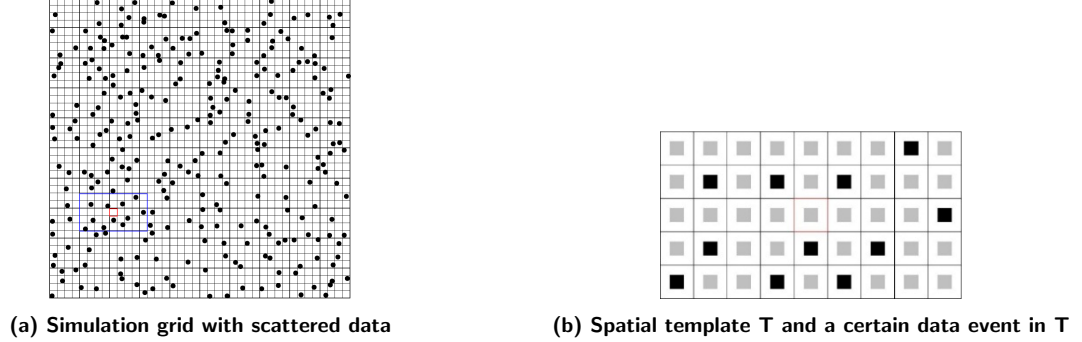


Figure 1: (a) A size 40×40 grid to be simulated with a 9×5 template overlaid on the current visiting node; (b) Spatial template $\mathbf{T}$ and a certain data event in $\mathbf{T}$, the center square is the node to be simulated; the black squares are the conditioning data.

The spatial moments of a random field $\mathbf{Z}$ in a template $\mathbf{T}$ can be expressed element-wise as

$$\text{Mom}_{\mathbf{Z}}^{\mathbf{T}}(w_0, \dots, w_N) = E[\mathbf{h}_1, \dots, \mathbf{h}_N; Z_0^{w_0} \dots Z_N^{w_N}], \quad (4)$$

where $\text{Mom}_{\mathbf{Z}}^{\mathbf{T}}$ is the moment function of $\mathbf{Z}$ in the spatial template $\mathbf{T}$, $(\mathbf{h}_1, \dots, \mathbf{h}_N)$ are the distance vectors to represent the geometry of $\mathbf{T}$ and w_i are the orders of the moments with each random variable $Z(\mathbf{u}_i)$ ($i = 1, \dots, N$).

The Legendre polynomials are used here to further define the concept of spatial Legendre moments. Legendre polynomials are one kind of special math functions defined on the interval $[-1, 1]$, which can be expressed using Rodrigues' formula (Zarowski 2004)

$$P_m(z) = \frac{1}{2^m m!} \frac{d^m}{dz^m} [(z^2 - 1)^m], \quad (5)$$

where $P_m(z)$ is the m th-degree Legendre polynomial. The infinite sequence of polynomials forms a complete orthogonal basis set on the domain $D = [-1, 1]$. The orthogonal property of the Legendre polynomials can

be expressed as

$$\int_D P_m(z) P_n(z) dz = \begin{cases} 0 & m \neq n \\ \frac{2}{2m+1} & m = n \end{cases} \quad (6)$$

and the norm of the Legendre polynomial $P_m(z)$ is

$$\|P_m\| = \sqrt{\frac{2}{2m+1}}. \quad (7)$$

With a simple substitution of polynomials in moment function (4) into Legendre polynomials, the spatial Legendre moments are defined as

$$L_{w_0 w_1 \dots w_N}^{\mathbf{T}} = \prod_{i=0}^N \left(w_i + \frac{1}{2} \right) \cdot E[h_1, \dots, h_N; P_{w_0}(z_0) P_{w_1}(z_1) \dots P_{w_N}(z_N)], \quad (8)$$

where $L_{w_0 w_1 \dots w_N}^{\mathbf{T}}$ are Legendre moments defined on the spatial template $\mathbf{T}$; the extra coefficient $(w_i + \frac{1}{2})$ on the right-hand side of the equation is intentionally introduced as a normalization term for convenience of the later computation (see Appendix for the details).

2.3 Multivariate expansion series of joint probability density function

A piecewise continuous function $f(z)$ defined on the interval $[-1, 1]$ can be written as a series of Legendre polynomials

$$f(z) = \sum_{m=0}^{\infty} L_m P_m(z). \quad (9)$$

Likewise, the expansion of a multivariate function $f(z_0, z_1, \dots, z_N)$ can be defined on an $N + 1$ -dimensional domain in the same way. Specifically, suppose that the multivariate function is a density function related to the joint distribution of random variables on a spatial template $\mathbf{T}$. The density function can be expanded into Legendre polynomial series in terms of Legendre spatial moments and Legendre polynomials as (see Appendix for the details)

$$f(z_0, z_1, \dots, z_N) = \sum_{w_0=0}^{\infty} \sum_{w_1=0}^{\infty} \dots \sum_{w_N=0}^{\infty} L_{w_0 w_1 \dots w_N}^{\mathbf{T}} P_{w_0}(z_0) P_{w_1}(z_1) \dots P_{w_N}(z_N). \quad (10)$$

In practice, the above infinite series, Equation (10), are truncated at a certain order W , thus leading to the approximated density function

$$f(z_0, z_1, \dots, z_N) \approx f_W(z_0, z_1, \dots, z_N) = \sum_{w_0=0}^W \sum_{w_1=0}^W \dots \sum_{w_N=0}^W L_{w_0 w_1 \dots w_N}^{\mathbf{T}} \prod_{i=0}^N P_{w_i}(z_i). \quad (11)$$

From definition (8), the spatial Legendre moments can be explicitly derived as

$$L_{w_0 w_1 \dots w_N}^{\mathbf{T}} = \int_D \prod_{i=0}^N \left[\left(w_i + \frac{1}{2} \right) P_{w_i}(z_i) \right] f(z_0, z_1, \dots, z_N) dz_0 dz_1 \dots dz_N. \quad (12)$$

Experimentally, if there are M replicates of data events associated with template $\mathbf{T}$ found in the training image, the spatial Legendre moments can be calculated as

$$\tilde{L}_{w_0 w_1 \dots w_N}^{\mathbf{T}} = \frac{1}{M} \sum_{t=1}^M \prod_{i=0}^N \left(w_i + \frac{1}{2} \right) P_{w_i}(\zeta_{t,i}), \quad (13)$$

where $\zeta_{t,i}$ are the data values of replicates in the template $\mathbf{T}$, t is the sequence number of replicates and i is sequence number of random variables.

3 Computational model

Combining the Equations (10) to (13), the empirical joint *pdf* can be derived as

$$\begin{aligned}\tilde{f}(z_0, z_1, \dots, z_N) &\approx \tilde{f}_W(z_0, z_1, \dots, z_N) = \frac{1}{M} \sum_{t=1}^M \sum_{w_0=0}^W \sum_{w_1=0}^W \cdots \prod_{i=0}^N \left[\left(w_i + \frac{1}{2} \right) P_{w_i}(\zeta_{t,i}) P_{w_i}(z_i) \right] \\ &= \frac{1}{M} \sum_{t=1}^M \prod_{i=0}^N \left[\sum_{w=0}^W \left(w + \frac{1}{2} \right) P_w(\zeta_{t,i}) P_w(z_i) \right]\end{aligned}\quad (14)$$

Equation (14) gives a unified computational model of empirical estimation of the density function on the spatial template $\mathbf{T}$, noticing that, in the right-hand side of the equation, the subscript i of w_i is dropped because of the symmetry of computation.

Now let's consider the *cpdf* $f_{Z_0}(z_0 | \Lambda)$ of a single sampling step in sequential simulation (ref. Section 2.1). The joint *pdf* can be marginalized from the Equation (14) to get the marginal *pdf* of conditioning random variables. To specify the difference between the empirical models and theoretical models in Equations (10)–(11), $\tilde{f}$ and $\tilde{f}_W$ specifically denote the experimental function corresponding to probability density function f and its Legendre polynomial series truncated at order W , respectively.

For convenience, denote functions $X_t(z_i)$ as

$$X_t(z_i) = \sum_{w=0}^W \left(w + \frac{1}{2} \right) P_w(\zeta_{t,i}) P_w(z_i). \quad (15)$$

Then, Equation (14) can be rewritten as

$$\tilde{f}_W(z_0, z_1, \dots, z_N) = \frac{1}{M} \sum_{t=1}^M X_t(z_0) \prod_{i=1}^N X_t(z_i). \quad (16)$$

The result of integration of $X_t(z)$ over $[-1, 1]$ can be derived from the orthogonal properties of Legendre polynomials as

$$\int_{-1}^1 X_t(z_i) dz_i = 1. \quad (17)$$

In fact, Equations (16) and (17) ensures that the integral of the approximated probability density function to be 1, a necessary property of probability density.

Followed by the marginalization and Equation (17), the empirical density of marginal distribution on the random variables $z_1, \dots, z_N$ is

$$\tilde{f}_W(z_1, \dots, z_N) = \frac{1}{M} \sum_{t=1}^M \prod_{i=1}^N X_t(z_i). \quad (18)$$

From Equations (16) and (18) and considering the relation between the conditional probability density function and the joint probability density function, one can derive

$$f(z_0 | \Lambda) \approx \tilde{f}_W(z_0 | \Lambda) = \frac{\sum_{t=1}^M X_t(z_0) \cdot \prod_{i=1}^N X_t(\zeta_i)}{\sum_{t=1}^M \prod_{i=1}^N X_t(\zeta_i)}, \quad (19)$$

which provides a concise computational model of the *cpdf*.

The above development provides a theoretical equivalency of the approximation of *cpdf* by truncated Legendre series, which was proposed in Mustapha and Dimitrakopoulos (2010b, 2011). However, the new reformulated model in the current paper leads to a different stochastic simulation method in view of the

related computational aspects. The advantage of the new model represented by Equation (19) is that no explicit computations of moments or cumulants are needed. In addition, the new model is computationally more accurate than the *hosim* program in Mustapha and Dimitrakopoulos (2011), in which some terms have to be dropped from the full expansion of Legendre series in the form of spatial cumulants to gain computational efficiency.

4 Algorithm description and computational analysis

4.1 Algorithm for computing cpdf

From the Equations (17) to (19), it can be easily shown that

$$\int_{-1}^1 \tilde{f}_W(z_0 | \Lambda) dz_0 = 1. \quad (20)$$

As $X_t(\zeta_{t,i})$ is a constant from the Equation (15), and from Equations (15) and (19), it is obvious that $\tilde{f}_W(z_0 | \Lambda)$ can be expressed as the summation of a series of Legendre polynomials, that is

$$\tilde{f}_W(z_0 | \Lambda) = \sum_{w=0}^W c_w P_w(z_0), \quad (21)$$

where $c_w (w = 1, \dots, W)$ are constants which can be conveniently computed as shown in the following Algorithm 1.

By the property of Legendre polynomial that $P_0(z) = 1, \forall z \in [-1, 1]$, combined with Equations (15) and (21), the computation of coefficients $c_w (w = 1, \dots, W)$ can be divided into the computation of functions $X_t(z_i)$ over the nodes of each replicate. Especially, the first term of c_w is always fixed as $c_0 = \frac{1}{2}$.

4.2 Recursive algorithm for computing a ccdf

From the results of Algorithm 1, the conditional probability density function can be expressed as

$$f(z_0 | \Lambda) = \frac{1}{2} + \sum_{w=1}^W c_w P_w(z_0). \quad (22)$$

The coefficient $c_0 = \frac{1}{2}$ is taken out from the summation in Equation (22) so that the Bonnet's recursion relation of Legendre polynomials can be smoothly applied in the followed derivation.

According to the Bonnet's recursion relation of Legendre polynomials

$$(2w+1)P_w(z) = \frac{d}{dz} [P_{w+1}(z) - P_{w-1}(z)], \quad (23)$$

the following equation can be derived

$$(2w+1) \int_{-1}^{z_0} P_w(z) dz = P_{w+1}(z_0) - P_{w-1}(z_0). \quad (24)$$

Therefore, the conditional cumulative distribution function (*ccdf*), $F(z_0 | \Lambda)$, can be deduced as

$$\begin{aligned} F(z_0 | \Lambda) &= \int_{-1}^{z_0} f(z_0 | \Lambda) dz \\ &= \frac{1}{2} + \frac{1}{2} z_0 + \sum_{w=1}^W \frac{c_w}{2w+1} [P_{w+1}(z_0) - P_{w-1}(z_0)], \\ &= \sum_{w=0}^{W+1} d_w P_w(z_0) \end{aligned} \quad (25)$$

Algorithm 1 Calculation of a *cpdf*

Data:
(1) data event (conditioning data): $\Lambda = \{\zeta_1, \dots, \zeta_N\}$;
(2) replicates of data events: $\zeta_{t,i}$, $t = 1, \dots, M$; $i = 1, \dots, N$;
(3) maximum order: W
Result: conditional probability density function
// Initialize the coefficients
for $w = 0$ **to** W **do**
 $c[w] = 0$;
end for
for $t = 1$ **to** M **do**
 //Computation of the function $X_t(z_0)$
 for $w = 0$ **to** W **do**
 $X[w] = (w + \frac{1}{2}) \cdot P_w(\zeta_{t,0})$;
 end for
 //According to Equation (19), the product $\prod_{i=1}^N X_t(\zeta_i)$ needs to be computed.
 //The product is initialized as 1 before the calculation of $X_t(\zeta_i)$
 $X_Prod = 1$;
 // Update the product $\prod_{i=1}^N X_t(\zeta_i)$ by computing the function $X_t(\zeta_i)$
 for $i = 1$ **to** N **do**
 $X_t = 0$;
 for $w = 0$ **to** W **do**
 $X_t = X_t + (w + \frac{1}{2}) \cdot P_w(\zeta_{t,i}) \cdot P_w(\zeta_i)$
 end for
 $X_Prod = X_Prod \cdot X_t$;
 end for
 for $w = 0$ **to** W **do**
 $X[w] = X[w] \cdot X_Prod$;
 $c[w] += X[w]$;
 end for
end for
//Note that the denominator in Equation (19) $\sum_{t=1}^M \prod_{i=1}^N X_t(\zeta_i)$ is actually equal to $2 \cdot c[0]$
 $denom = 2 \cdot c[0]$;
for $w = 0$ **to** W **do**
 $c[w] = c[w] / denom$;
end for
END OF ALGORITHM

As can be seen from Equation (25), the *ccdf* is also expressed as the summation of the univariate Legendre polynomials, with the order of the Legendre polynomials increasing by one because of the integration. Furthermore, the new coefficients d_w ($w = 0, \dots, W, W + 1$) now can be computed through Equation (25) in an iterative way, as shown in Algorithm 2.

Algorithm 2 Integration of a *cpdf* to get a *ccdf*

Data:
(1) maximum order of *cpdf*: W ;
(2) coefficients array of *cpdf*: $c[0, 1, \dots, W]$.
Result: coefficients array of cumulative distribution function: $d[0, 1, \dots, W + 1]$.
// Initialize the coefficients
for $w = 0$ **to** $W + 1$ **do**
 $d[w] = 0$;
end for
// Update the coefficients according to Equation (25)
for $w = 1$ **to** W **do**
 $d[w + 1] = d[w + 1] + \frac{c[w]}{2w + 1}$;
 $d[w - 1] = d[w - 1] - \frac{c[w]}{2w + 1}$;
end for
 $d[0] = d[0] + \frac{1}{2}$;
 $d[1] = d[1] + \frac{1}{2}$;
END OF ALGORITHM

4.3 Computational complexity

The most computationally demanding part of the high-order simulation algorithm is to calculate the Legendre series coefficients, which is the basis for estimating the conditional probability density functions. Considering that the conditional probability density functions are approximated by Legendre series truncated to a certain order W , as Equation (11) shows, the number of the different coefficients is $(W + 1)^{N+1}$, where N is the number of data points. Even the Legendre series are approximated by truncated series, where the sum of orders of different variables is not greater than W , which is the form adopted in Mustapha and Dimitrakopoulos (2011). The number of the different coefficients is still as big as $\sum_{w=0}^W \binom{N+w}{w}$ for a single data event. Although this computational complexity can be reduced by discarding some terms which are regarded as negligible, it should be noted that this simplification may lead to a loss of accuracy.

From Equations (15) and (19), it can be seen that all of the different coefficients introduced by the explicit expansion of Legendre series are reduced to a calculation of the function $\prod_{i=1}^N X_t(z_i)$. There are only NW computations of Legendre polynomials and a few products and additions included in the calculation of the function $\prod_{i=1}^N X_t(\zeta_{t,i})$ for each data event encountered in the training image. It should be noted that the computational time still depends on the number of the replicates encountered in the training image, as well as the maximal order of Legendre polynomials and the number of conditioning in the neighborhood. However, the computational cost regarding the above-mentioned parameters is significantly reduced, as opposed to computing the large number of coefficients in the previous version of high-order simulation.

5 Implementation

The implementation is relatively straightforward in terms of the above algorithms estimating the *cpdf* and *ccdf* according to the framework of sequential simulation. However, a method is proposed in this section to deal with the replicates, aiming to reduce the conflicts of spatial statistics between the sample data and the training image. The main idea of the method is to deliberately select replicates which are similar to the conditioning data within a certain range according to some measure of similarity. The reason is that the conditional probability distribution is a one-dimensional intercept from the multivariate joint probability distribution, and therefore the replicates that are close to the conditioning data are more relevant to estimate this one-dimensional local probability distribution.

For every node to be simulated in sequential simulation, a local neighborhood is defined to search for conditional data from both the sample data and the simulation grid. The locations of these conditional data together with the center node to be simulated constitute a geometry template. Given a training image, replicates of the geometry template can always be found from the training image as long as the searching neighborhood is inside of the training image's extent. In the present work, the measure of similarity between the replicates and the data event is set to be the average square Euclidean distance between the replicates and the conditioning data, and the threshold is set as the variance of the sample data. The replicate will be selected in the estimation of *cpdf* if the distance between the replicate and the conditioning data is less than the variance of the sample data. In addition, when there are few replicates that can be found from the training image due to the conflicts between the sample and the training image, some tolerances are given to the shape of the geometry template so that similar replicates can be found. Figure 2 shows a general way to search the candidate points associated with a certain vector in a spatial template. The parameter θ is the angle tolerance of the candidate point's deviation from the original vector in the template, and Δh and b are the tolerances in the lag and bandwidth, respectively. Possible candidate points are taken from the shadowed area, and the point that has the closest property to the ending node of the original vector in the template is selected. To maintain the consistency of the geometry configuration, an inner part of the template is specified such that the relative locations to the center node inside the inner part remains unchanged. In other words, only the nodes further away from the center node are allowed to have the ability to change locations. This strategy gives more flexibility to manipulate the geometry configuration of the replicates.

The main procedure of the high-order simulation approach can be summarized in the following steps:

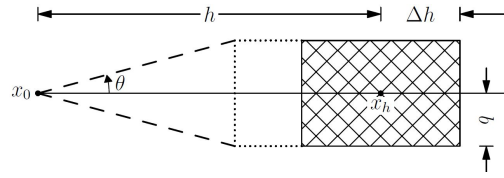


Figure 2: Finding approximate replicates from the training image with the tolerances of the original geometry template.

1. Read the sample data and training image into memory. In order to apply the multi-variate expansion of Legendre polynomials, the property values of the samples or training image are scaled to the interval $[-1, 1]$ through a linear transformation.
2. Specify dimensions of a certain neighborhood for searching the conditional data and other parameters such as the minimum or maximum number of the conditional data. The geometry of the local template totally depends on the locations of the conditional data. In the present work, a rectangular shape neighborhood was used and searching policy was applied to find the closest points to the center. Nevertheless, the shape of the neighborhood and the searching policy can be manipulated to further control the spatial configuration of the template.
3. Set the lag tolerance, angle tolerance and bandwidth tolerance to enable searching approximate replicates from the training image (see Figure 2).
4. Generate a random sequence on the indices of the simulation grid to create a random visiting path.
5. According to the predefined visiting path, sequentially pick one node at a time for the simulation. If the property value is already known (copied from the hard data), then continue to choose another single node until the property value is not assigned. The conditioning data are searched inside the neighborhood centered on the chosen node by the previously specified searching policy from both the hard data and the simulated nodes.
6. A local spatial template is determined by the data and the center node for later simulation. This spatial template is then used to find similar replicates from the training image according to the parameters set in Steps 2 and 3. If the number of approximated replicates is not adequate for statistical inference, then drop the furthest node to the center node and repeat until the minimum number of conditioning data is reached.
7. The local conditional cumulative distribution function is estimated from the replicates using the algorithms elaborated in Section 3. A random value is drawn from the local *cdf* using the Monte Carlo method and set as the property value of the node to be simulated.
8. Repeat from Step 5 until all the nodes in the random path are visited.

6 Examples and comparisons

The data used in this paper are extracted from the Stanford V reservoir data set (Mao and Journel 1999). A horizontal section serving as the exhaustive image is taken from the Stanford V reservoir model of porosity in a square grid with 100×100 pixels (cells of size 100 meters). As seen from the exhausting image in Figure 3, porosity values are distributed as several channels that can be distinguished from the background. For the examples and comparisons presented in the next sections, 200 data points are randomly sampled from the selected exhaustive image to serve as the sample data set and are displayed in Figure 4. Applying the proposed high-order approach, the selected data is used to simulate the exhaustive image in two different ways, so as to show the sensitivity of the approach to the chosen TI. Accordingly, in Example 1, the exhaustive data is used as the TI; then, in Example 2, the TI is selected from a different section of the Stanford V reservoir data set than the exhaustive image. The second TI is shown in Figure 5 and has different spatial patterns than those in the exhaustive image. In addition, a comparison of the proposed algorithm to the well-established *mps* method *filtersim* (Zhang et al. 2006) is presented. In each of the realizations using the high-order simulation algorithm, a window of size 15×20 in terms of cell size is used as the search template. The tolerance angle for searching is set to 15 degrees, the lag tolerance to 2 and the bandwidth to 1. These parameters are chosen

from the calculation of experimental variograms (Goovaerts 1997). The minimum number of conditioning data is 6 and the maximum number is 12, while 5 previously simulated values are used. The maximum order of Legendre polynomials is set to 10. For the realizations generated with *filtersim*, the searching template is 15×21 with an inner patch of size 7×7 and a multiple grid level of 3, while replicates are classified into different categories according to their filter scores. For further details on *filtersim*, the reader is referred to Zhang et al. (2006).

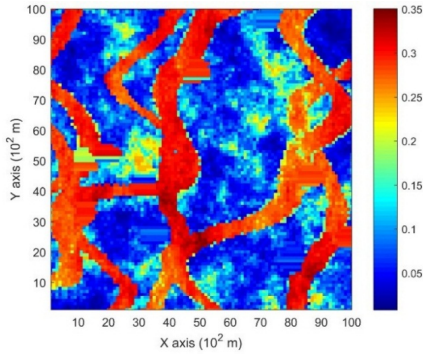


Figure 3: A horizontal section from reservoir's porosity values with sinuous connectivity.

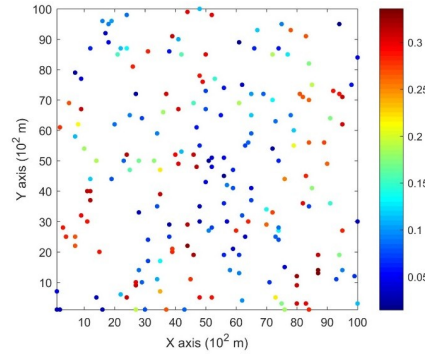


Figure 4: Data points sampled from the exhaustive image (containing 200 points, or 2% of the total data).

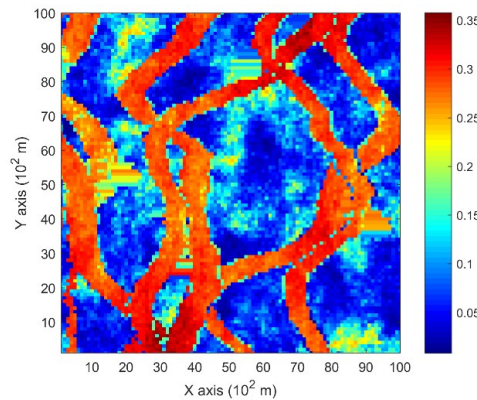


Figure 5: Training image that is different from the exhaustive data.

6.1 Example 1

This example generates simulations using the 200 samples shown in Figure 4 and the exhaustive image in Figure 3 as the TI. In this case, there are no conflicts between the available data and the training image. Figure 6 shows one realization from the high-order simulation and another from *filtersim*, respectively. From visual comparison with the exhaustive image, the realization from the high-order simulation reproduces better the channels of the original image. To demonstrate the reproduction of the distribution and second-order spatial statistics of simulation results, 10 different realizations for each method are generated. The histograms of the realizations are displayed in the Figure 7 and related variograms are displayed in Figure 8. Both simulation methods reproduce well the bimodal shape in the histograms; however, in general, high-order simulations show better reproduction in the proportions of porosity values. High-order simulation methods also reproduce well the variograms in the X-direction or Y-direction, while the variograms from the *filtersim* simulations demonstrate larger fluctuations and have notable deviations from the variogram of the exhaustive data in the Y-direction. For a comparison of the high-order spatial statistics of simulation results to the original data in the two different settings, the third-order cumulant maps are generated by the HOSC program (Mustapha and Dimitrakopoulos 2010c). This program uses a template with two directions in X-axis and Y-axis, and the number of lags are 70 with lag size as 1, which are displayed in Figure 9. In

comparison to the third-order cumulant map of the exhaustive image, the high-order simulation performs better in the reproduction of the high-order statistics, although both simulation methods have reasonable similarity in terms of the third-order cumulant map, as there are no conflicts between the sample data and the training image in this case.

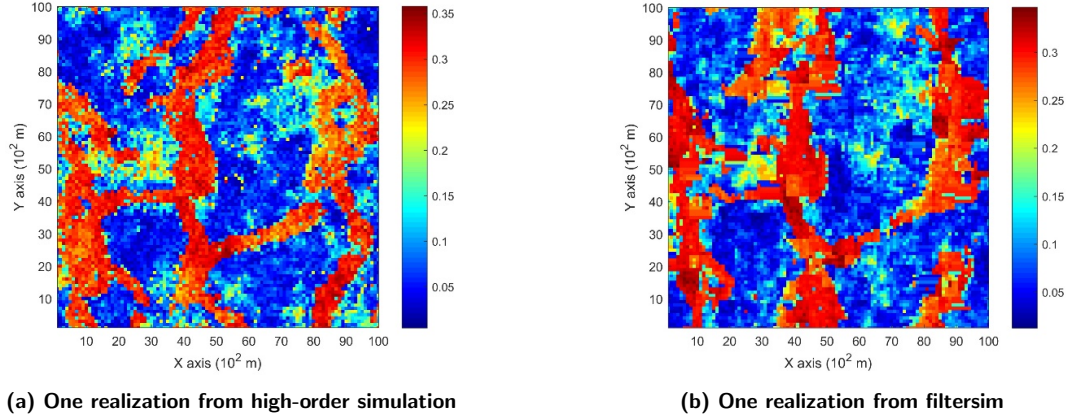


Figure 6: Simulations with 200 sample data using the exhaustive data as the training image. one realization from (a) high-order simulation and (b) filtersim.

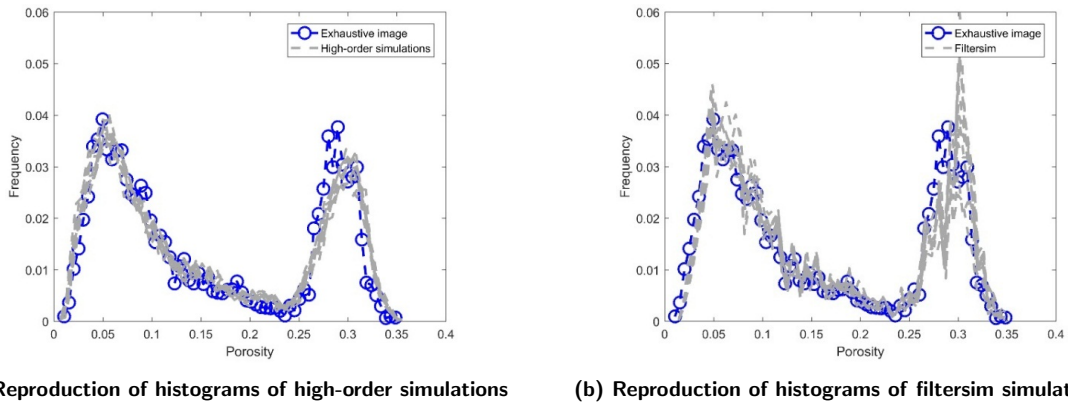


Figure 7: Reproduction of histograms of 10 realizations with 200 sample data using the exhaustive data as the training image: (a) and (b) correspond to 10 realizations from the high-order simulation and filtersim, respectively.

6.2 Example 2

In this setting, the simulations are conducted with the same conditioning data; however, the TI is different from the exhaustive data. Figure 10 shows one realization from the high-order simulation and one for *filtersim*. Clearly, there are conflicts between the spatial statistics of the sample data and the TI, which are key factors affecting the results of the simulations. As expected, the reproduction of the spatial patterns is worse when compared to the results from the simulations in the previous example. Nevertheless, the realization from high-order simulation method still maintains the spatial structures of the original exhaustive data. As shown in Figure 11, the 10 realizations of the high-order simulation match the histogram of the exhaustive image very well. By contrast, the 10 realizations of *filtersim* mismatched the exhaustive image in some part of the proportions. From the comparison shown in Figure 12, the high-order simulation performs better than *filtersim* in reproducing the variograms of the exhaustive image as well, although there is a minor deviation in the Y- direction.

In order to demonstrate the impact of the conflicts between the sample data and training image during the simulations, Figure 13 (a)–(e) shows the third-order cumulant maps corresponding to the exhaustive image, the sample data, the training image and one realization of high-order simulation and *filtersim*. The

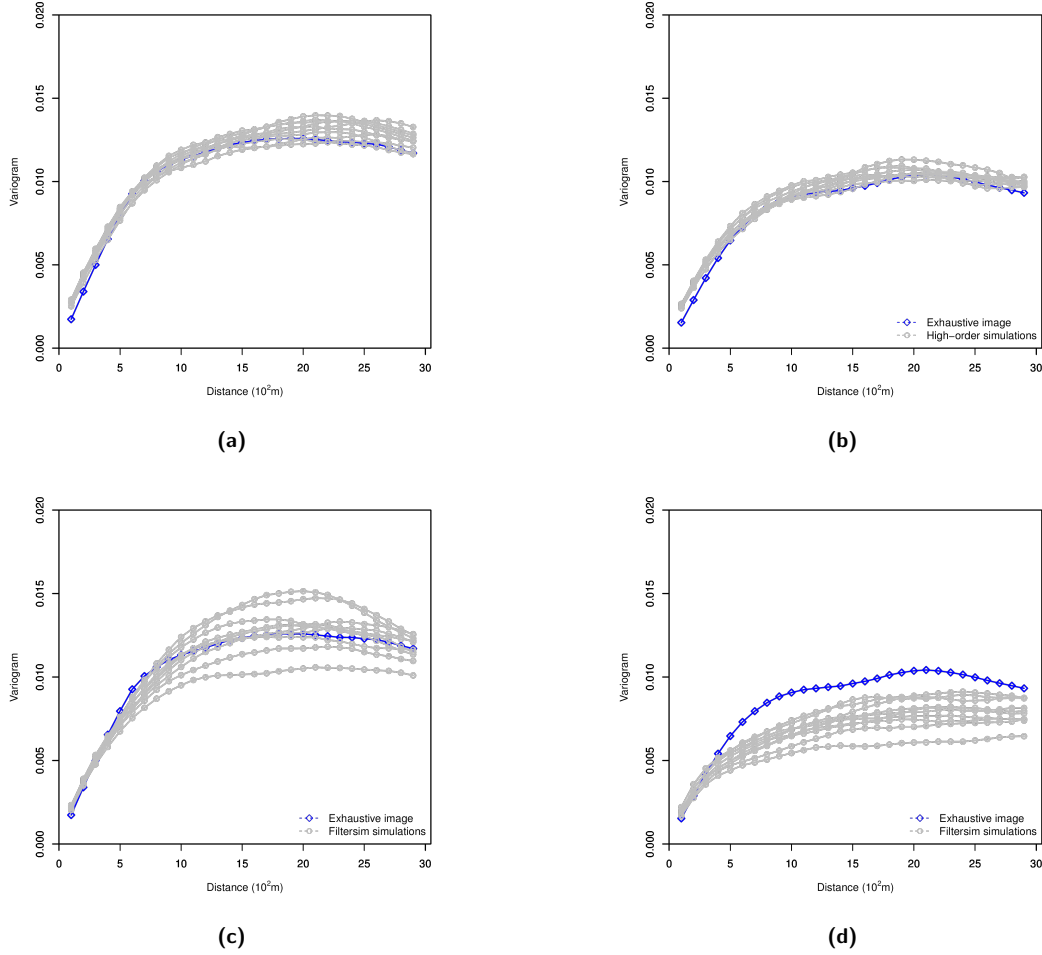


Figure 8: Reproduction of the variograms of 10 realizations with 200 sample data using the exhaustive data as the training image from high-order simulation and filtersim, respectively. (a) Reproduction of variograms of high-order simulations in the X direction; (b) Reproduction of variograms of high-order simulations in the Y-direction; (c) Reproduction of variograms of filtersim simulations in the X-direction; (d) Reproduction of variograms of filtersim simulations in the Y-direction.

parameter settings to generate the cumulant maps for the grid data are the same as used in Figure 9, whereas the lag size is set to 5 grid cells, with the lag tolerance being set to 1 grid cell and the angle tolerance being set to 15 degrees for generating the cumulant map of the sample data. As the sample data is too sparse to compute the cumulant map at the same scale as the exhaustive image, some smoothing has been applied to the cumulant map of the sample data for the purpose of visualization. The third-order cumulant map of the realization from the high-order simulation maintains the main structures of the exhaustive data. On the other hand, the third-order cumulant map of the realization from *filtersim* resembles the cumulant map of the TI, which differs from the cumulant map of the exhaustive image. This implies that the high-order simulation is primarily data-driven, whereas the *filtersim* method is TI-driven. This result can be explained by the fact that the high-order simulation seeks to find replicates that comply to the statistical configuration of the conditioning data from the training image, and the values of nodes to be simulated are drawn from the related local probability distribution. By contrast, the *filtersim* method is TI-driven, which means that the values of nodes to be simulated comes directly from the paste of certain replicates from the training image, which is patch-based instead of node-by-node as in the high-order simulation. In particular, the impact of the conditioning data is more important for capturing the large-scale spatial structures in the early stage of the high-order simulation. For instance, Figure 13 (b) shows the cumulant map of the sample data, and the resolution of the map is much coarser than the exhaustive data. This map shows some distortion when representing the third-order statistics of the exhaustive image due to the sparsity of the data. However, the

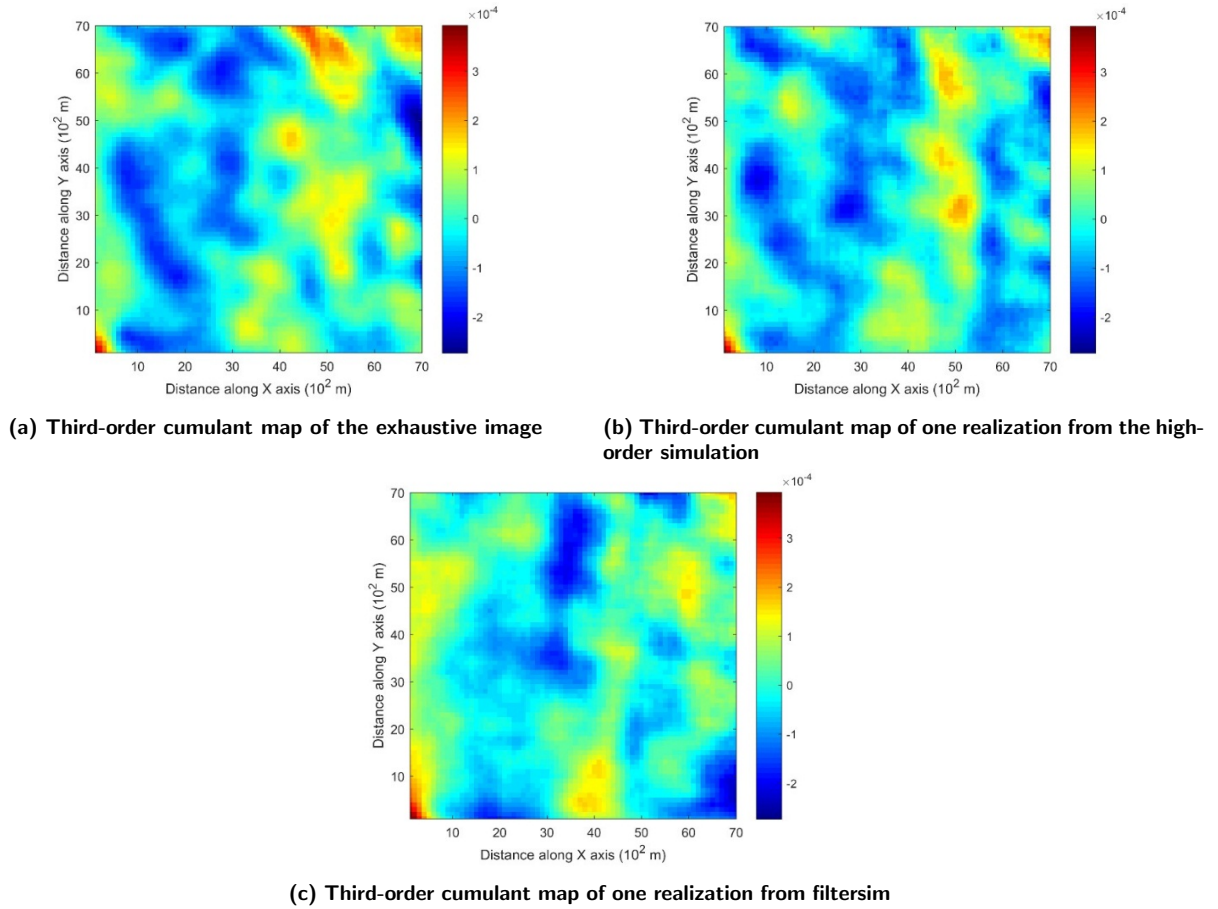


Figure 9: Comparing third-order cumulant maps of realizations with 200 sample data using the exhaustive data as the training image from the high-order simulation and filtersim, respectively.

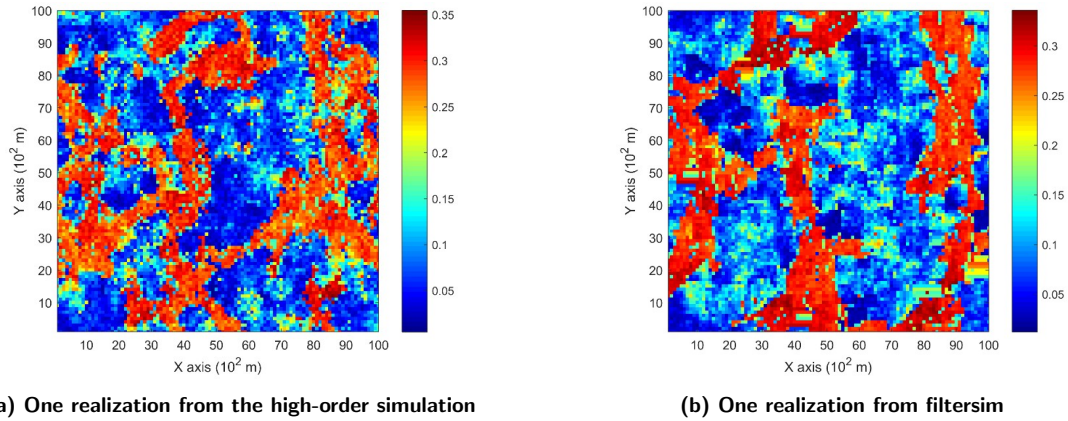


Figure 10: Simulations with 200 sample data using a separate training image different from the exhaustive data. one realization from (a) high-order simulation and (b) filtersim.

spatial structures of the limited sample data control the spatial statistics of the results from the high-order simulation. In general, the results in this case study show that the proposed high-order simulation algorithm can reasonably reproduce the overall probability distribution, the second-order statistics and the higher-order statistical features (such as spatial cumulants), as the statistical conflicts between the sample data and the training image are not severe.

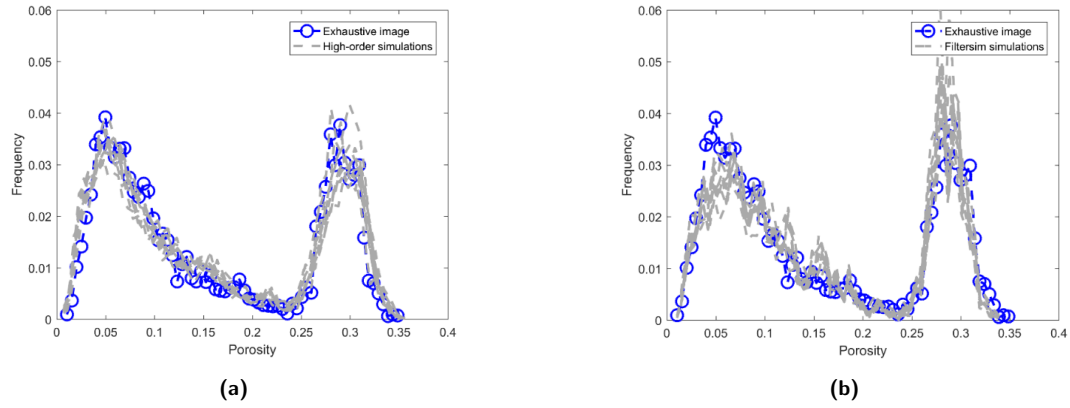
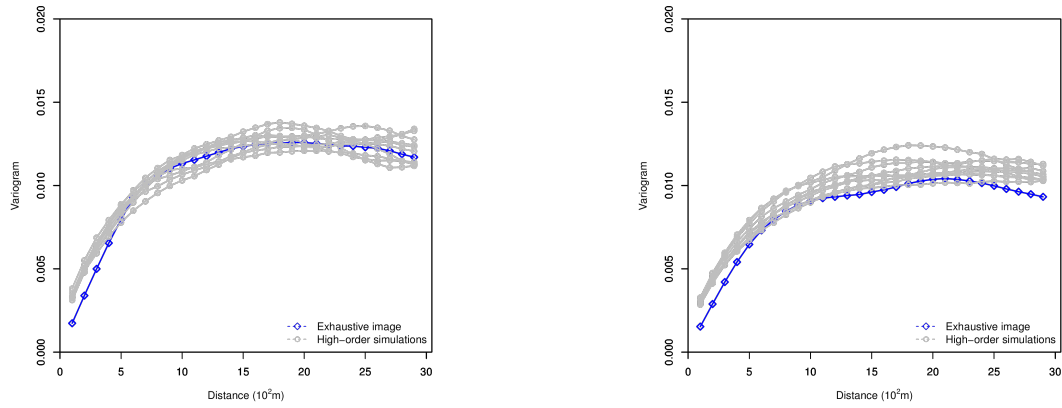
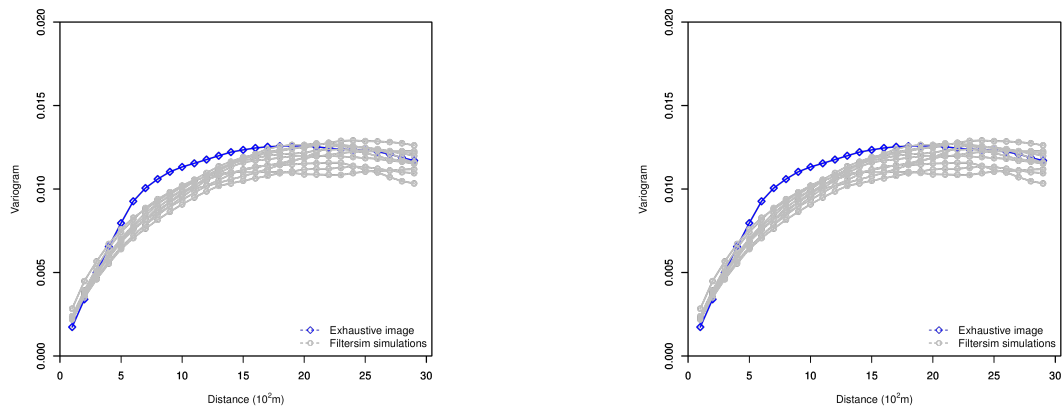


Figure 11: Reproduction of histograms of 10 realizations with 200 sample data using the training image different from the exhaustive data: (a) and (b) correspond to 10 realizations from high-order simulation and filtersim, respectively.



(a) Reproduction of variograms of high-order simulations in X direction

(b) Reproduction of variograms of high-order simulations in Y direction



(c) Reproduction of variograms of filtersim simulations in X direction

(d) Reproduction of variograms of filtersim simulations in Y direction

Figure 12: Variograms of 10 realizations with 200 sample data using the training image different from the exhaustive data from the high-order simulation and filtersim, respectively.

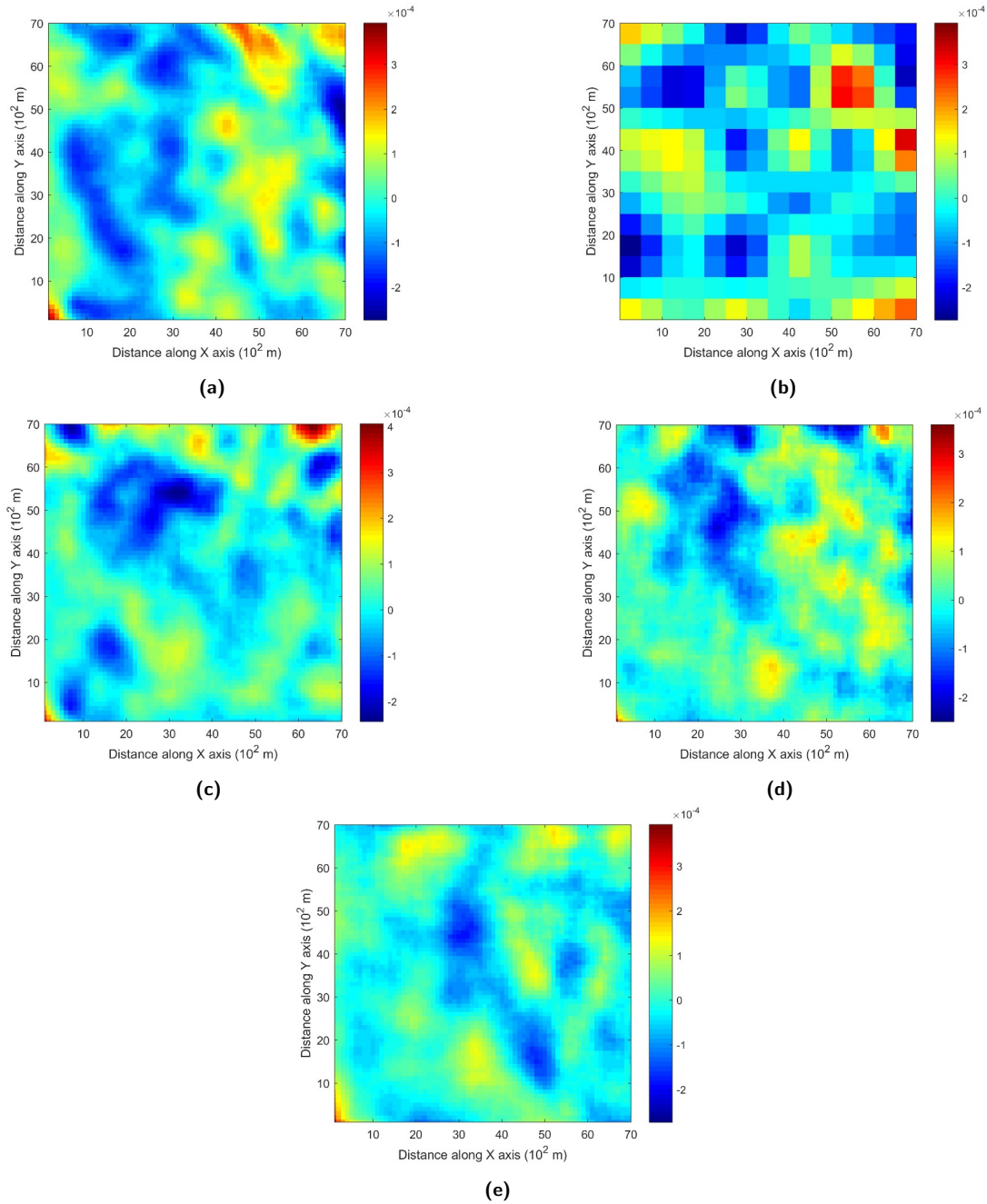


Figure 13: Third-order cumulant maps of (a) exhaustive image, (b) sample data, (c) TI, (d) high-order simulation and (e) filtersim.

6.3 Parameter sensitivity testing

Most parameters in the current implementation of high-order stochastic simulation method are experimental choices. Amongst all the parameters encountered in the current implementation, some follow common practices in the parameter selection for conventional geostatistical simulations, such as the size of the search window, the lag and angle tolerance. Additionally, in the high-order simulation method presented here, the number of conditioning data corresponding to a certain template needs more consideration, as it determines the dimension of the local probability distribution. In the current implementation, the number of the conditioning data is limited for two important reasons. First, the limited number of conditioning data reduces the computational time needed to estimate the *cpdf*. Second, the method resembles the so-called multiple

grid strategy (Strebelle 2002) applied in many multi-point simulation methods in order to maintain both large and small-scale spatial structures. In the early stage of the simulation process, the neighborhoods are more likely to capture large-scale patterns, since the known data are sparse. The neighborhoods gradually correspond to finer-scale patterns as the simulation continues and more known data are generated. A similar search strategy has also been applied and discussed in Mariethoz et al. (2010).

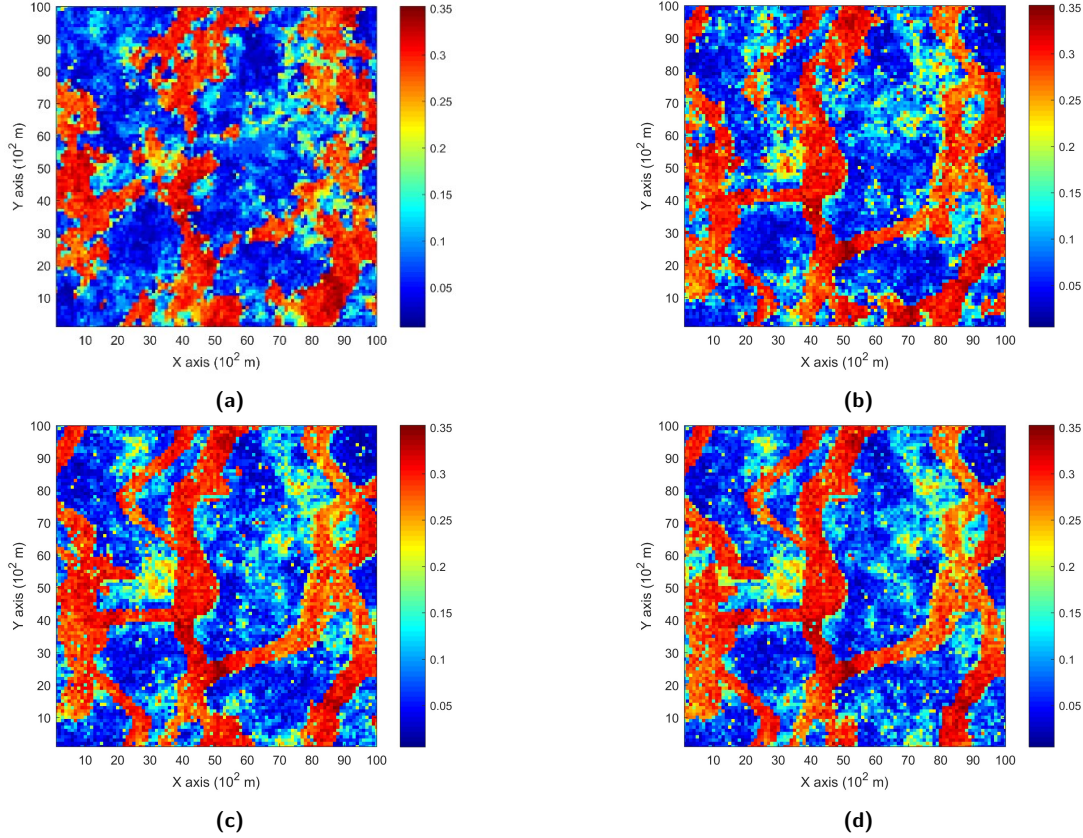


Figure 14: Comparing the realizations of high-order simulation by applying different local neighborhood size, with 200 sample data using the exhaustive data as the training image. The maximum order of Legendre polynomials to approximate the cpdfs is 10 for all the realizations. Realizations with neighborhood of (a) 6 conditioning data, (b) 12 conditioning data, (c) 20 conditioning data, and (d) 30 conditioning data.

The maximum order of the polynomials is another parameter of importance in the high-order simulation, since it affects the precision of the approximation of *cpdf* by a truncated Legendre polynomial series. Theoretically, the coefficients in the Legendre polynomial series decay exponentially, and in general much faster than in Taylor series (Cohen and Tan 2012; Wang and Xiang 2012). The numerical results in Cohen and Tan (2012) show that Legendre polynomial series with six non-zero coefficients (order 10 and 11 in their examples) are highly accurate approximations to the targets. The numerical testing to approximate a probability distribution regarding the order of Legendre polynomial series has also been investigated in Mustapha and Dimitrakopoulos (2010b) and led to similar results. However, it should be noted that the above tests are conducted for the approximation of a determined function, whereas for the approximation of the probability density function, there is also the impact from the limitation of the number of replicates. Depending on different data sets, Legendre polynomial series with order from 6 to 20 should be a reasonable range to select.

For validation and sensitivity analysis, further tests are conducted specifically to demonstrate the impacts of the number of conditioning data and the maximum order of Legendre polynomial series. In order to restrict the effects of the conflicting statistics between the TI and the sample data, the same data set from Example 1 is used to evaluate the sensitivity of the related parameters. The experiments are taken for each individual parameter without considering the possible dependencies between them. In all the experiments, the parameters not being tested remain the same as in Example 1. Furthermore, the random seed used to

generate the visiting path is also fixed for all the simulations in the experiments, so that the impact of the different visiting path is excluded. Figures 14 and 15 depict the realizations of the high-order simulation with different neighborhood sizes and their corresponding third-order cumulant maps. In addition, Figures 16 and 17 show the realizations of high-order simulation with respect to the order of the Legendre polynomial series in order to approximate the *cpdfs* (as well as their corresponding third-order cumulant maps). From the results, it can be seen that both the size of the neighborhood and the maximum order of the polynomials have considerable impacts on the high-order simulation results. In particular, using a small size of the neighborhood of 6 grid cells or a Legendre polynomial order less than 6 results in a poor reproduction of the spatial patterns as well as the cumulant maps. However, when the size of the neighborhood increases to more than 12 or the order of polynomials is greater than 10, the differences become trivial. Although the testing is for a specific data set, and the size neighborhood should be larger in 3D space than 2D space, it can be expected that a similar sensitivity analysis can be applied to choose the appropriate parameters on a case-by-case basis.

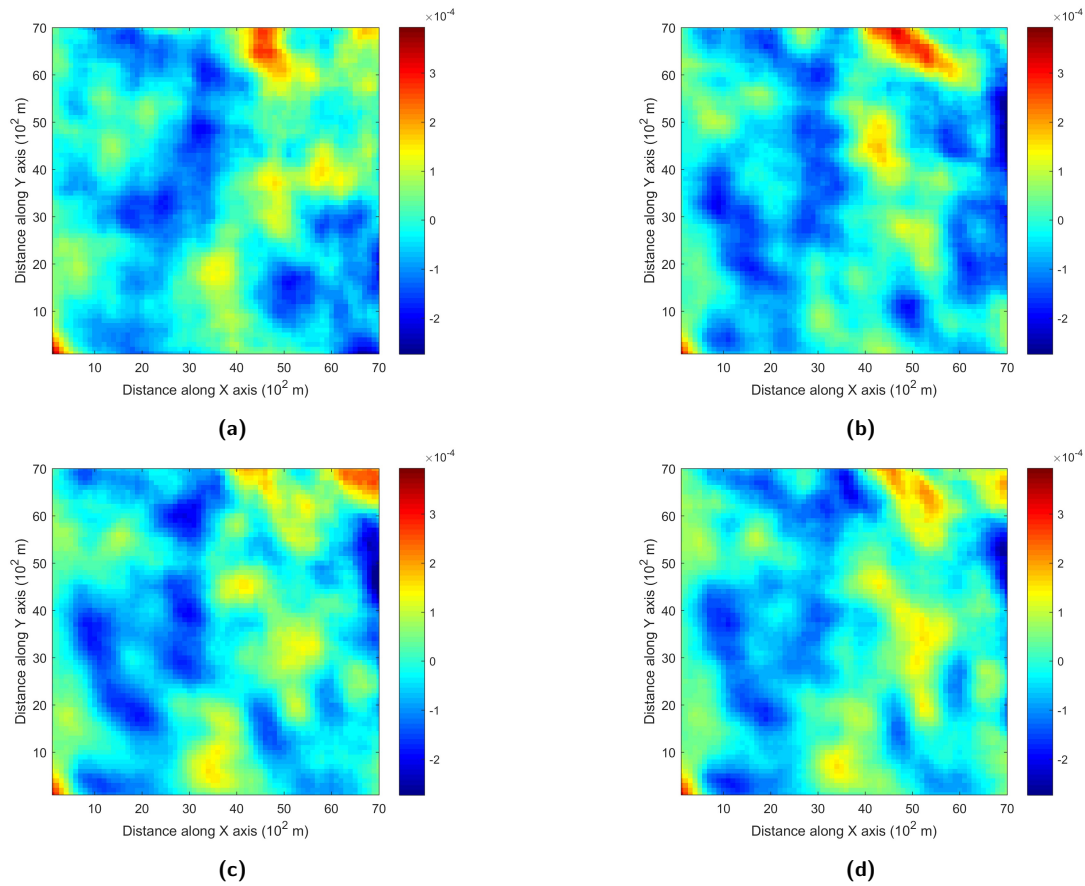


Figure 15: Comparing the third-order cumulant maps of the realizations of the high-order simulation by applying different local neighborhood size, with 200 sample data using the exhaustive data as the training image. The maximum order of Legendre polynomials to approximate the *cpdfs* is 10 for all the realizations. Third-order cumulant maps of one realization with neighborhood of (a) 6 conditioning data, (b) 12 conditioning data, (c) 20 conditioning data, and (d) 30 conditioning data.

7 Conclusions

The main contributions of the paper are as follows. Firstly, starting from the high-order simulation method based on Legendre polynomial series, a new computational model in the form of a unified empirical function is developed to approximate the *cpdf*. The computational model leads to an estimation of *cpdf* without calculating the high-order spatial cumulants or moments term by term. As a consequence, it not only greatly reduces the computational requirements but also provides a more accurate approximation of *cpdf* through

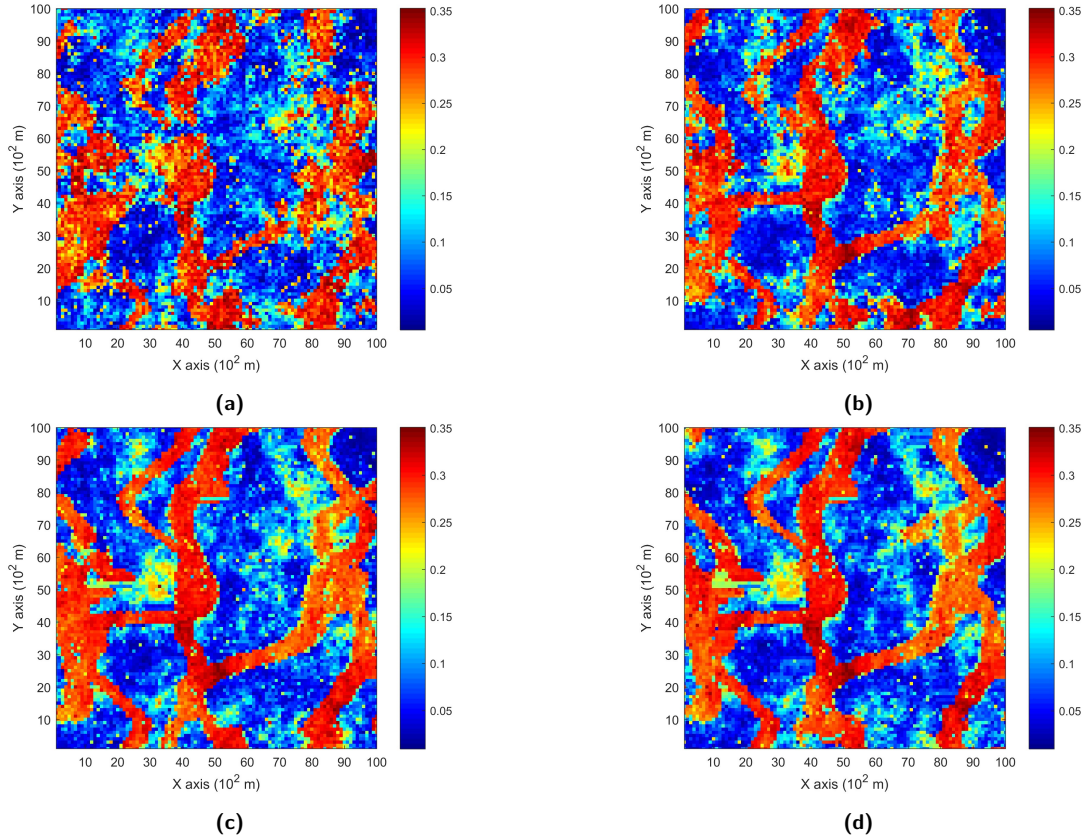


Figure 16: Comparing the realizations of the high-order simulation by applying the different order of truncated Legendre polynomial series, with 200 sample data using the exhaustive data as the training image. The number of conditioning data in the local neighborhood is 12 for all the realizations. Realizations of the high-order simulation by approximating *cpdf* with Legendre polynomial series up to (a) order 6, (b) order 10, (c) order 20, and (d) order 30.

Legendre polynomial series in comparison to the previous high-order simulation algorithm based on Legendre cumulants. Secondly, two new algorithms to derive *cpdf* and *ccdf* based on the above computational model are developed; they both use the properties of Legendre polynomials to simplify the computation and avoid an explicit expansion of a multivariate Legendre series. Lastly, the spatial template used in the current high-order simulation method is dynamically changing with the computation of the probability distribution in real time without storing data events. In addition, a flexible strategy to search replicates from the training image is proposed and implemented to deal with the conflicts between the statistics of the sample data and the training image.

Tests show the capacity of the proposed algorithm to reproduce complex geological patterns, and, in addition, that both the overall distribution and the high-order spatial statistics of the data are reproduced by the high-order simulations. Comparing the results of the high-order simulation in different cases with those of *filtersim*, the high-order simulation outperforms in the reproduction of high-order spatial statistics. This result becomes more notable in cases where there are conflicts in the spatial statistics between the sample data and the training image. This demonstrates that the high-order simulation has a more data-driven nature, whereas the *filtersim* is more TI-driven. Although the computational cost is significantly reduced (depending on the size of the training image, the number of neighborhood and the maximum order of Legendre polynomial series), the simulation is still slower than the *filtersim* method. However, since the computations of the *cpdf* is carried out on each replicate with the same type of calculation, the procedure could be parallelized so that the simulation can be further accelerated through parallelization techniques, such as GPU programming. It also should be noted that the approximation of *cpdfs* by Legendre series or any kind of polynomial series may generate problems of non-positive probability densities; further research is needed to address the issue.

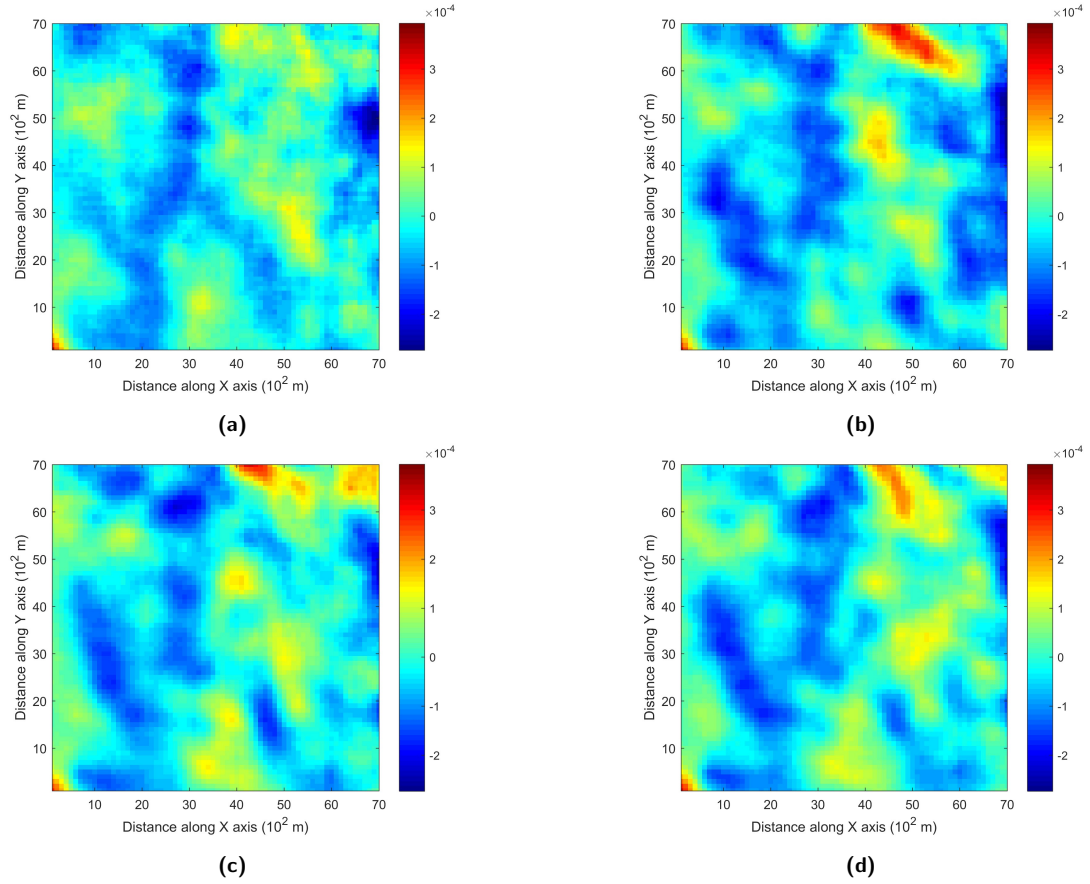


Figure 17: Comparing third-order cumulant maps of the realizations of the high-order simulation by applying the different order of truncated Legendre polynomial series, with 200 sample data using the exhaustive data as the training image. The number of conditioning data in the local neighborhood is 12 for all realizations. Third-order cumulant map of one realization of the high-order simulation by approximating cpdf with Legendre polynomial series up to (a) order 6, (b) order 10, (c) order 20, and (d) order 30.

8 Appendix: Expansion series of probability density function based on the spatial Legendre moments

Suppose that the multivariate function $f(z_0, z_1, \dots, z_N)$ is the density function related to the joint distribution of random variables on a spatial template $\mathbf{T}$, and that it can be expressed as a Legendre polynomial series. The sequence of Legendre polynomials at different orders forms a set of orthogonal bases of a Hilbert space containing all the continuous functions defined on $D = [-1, 1]^{N+1}$; the inner product is defined as

$$\langle g, h \rangle = \int_D gh \, dz_0 \dots dz_N, \quad (\text{A.1})$$

where g, h are functions in the Hilbert space.

From the orthogonal property of Legendre polynomial and the definition of its norm shown in Equation (7), there is:

$$f(z_0, z_1, \dots, z_N) = \sum_{w_0=0}^{\infty} \sum_{w_1=0}^{\infty} \dots \sum_{w_N=0}^{\infty} \langle f, \bar{P}_{w_0} \bar{P}_{w_1} \dots \bar{P}_{w_N} \rangle \bar{P}_{w_0} \bar{P}_{w_1} \dots \bar{P}_{w_N}, \quad (\text{A.2})$$

where the set $\{\bar{P}_{w_0} \bar{P}_{w_1} \dots \bar{P}_{w_N} | w_i = 0, 1, 2, \dots, 0 \leq i \leq N\}$ are the orthonormal bases of the Hilbert space, and $\bar{P}_{w_i}(z_i) = \frac{P_{w_i}(z_i)}{\|P_{w_i}\|}$, $0 \leq i \leq N$, is the normalized Legendre polynomial. Therefore,

$$\bar{P}_{w_0} \bar{P}_{w_1} \dots \bar{P}_{w_N} = \frac{\bar{P}_{w_0} \bar{P}_{w_1} \dots \bar{P}_{w_N}}{\|P_{w_0}\| \dots \|P_{w_N}\|} = \prod_{i=0}^N \sqrt{w_i + \frac{1}{2}} \cdot P_{w_0} P_{w_1} \dots P_{w_N} \quad (\text{A.3})$$

$$\langle f, \overline{P}_{w_0} \overline{P}_{w_1} \cdots \overline{P}_{w_N} \rangle = \prod_{i=0}^N \sqrt{w_i + \frac{1}{2}} \cdot \langle f, P_{w_0} P_{w_1} \cdots P_{w_N} \rangle. \quad (\text{A.4})$$

Combining Equations (A.2)–(A.4), it is:

$$f(z_0, z_1, \dots, z_N) = \sum_{w_0=0}^{\infty} \sum_{w_1=0}^{\infty} \cdots \sum_{w_N=0}^{\infty} \prod_{i=0}^N \left(w_i + \frac{1}{2} \right) \cdot \langle f, P_{w_0} P_{w_1} \cdots P_{w_N} \rangle P_{w_0} P_{w_1} \cdots P_{w_N}. \quad (\text{A.5})$$

Note that $f(z_0, z_1, \dots, z_N)$ is the probability density function, thus

$$\begin{aligned} \langle f, P_{w_0} P_{w_1} \cdots P_{w_N} \rangle &= \int_D P_{w_0}(z_0) P_{w_1}(z_1) \cdots P_{w_N}(z_N) f(z_0, z_1, \dots, z_N) dz_0 \cdots dz_N \\ &= E[h_1, \dots, h_N; P_{w_0}(z_0) P_{w_1}(z_1) \cdots P_{w_N}(z_N)]. \end{aligned} \quad (\text{A.6})$$

To use the Legendre polynomials as the bases without normalization and avoid computation of the square roots, the spatial Legendre moments are defined as

$$L_{w_0 w_1 \cdots w_N}^T = \prod_{i=0}^N \left(w_i + \frac{1}{2} \right) \cdot \langle f, P_{w_0} P_{w_1} \cdots P_{w_N} \rangle, \quad (\text{A.7})$$

which is equivalent to the definition in Equation (8).

Furthermore, from Equations (A.5)–(A.7), one can directly derive the expansion series of the probability density function based on the spatial Legendre moments, which appears in Equation (10). A similar derivation works for the truncated Legendre polynomial series, since the corresponding function space forms a finite-dimensional subspace of the above Hilbert space.

References

- Arpat GB, Caers J (2007) Conditional simulation with patterns. *Math Geol* 39(2):177–203. DOI:10.1007/s11004-006-9075-3
- Boucher A (2009) Considering complex training images with search tree partitioning. *Computers & Geosciences* 35(6):1151–1158. DOI:10.1016/j.cageo.2008.03.011
- Boucher A, Costa J, Rasera L, Motta E (2014) Simulation of geological contacts from interpreted geological model using multiple-point statistics. *Math Geosci* 46(5):1–12. DOI:10.1007/s11004-013-9510-1
- Chatterjee S, Mustapha H, Dimitrakopoulos R (2016) Fast wavelet-based stochastic simulation using training images. *Computational Geosciences* 20(3):399–420. DOI:10.1007/s10596-015-9482-y
- Cohen MA, Tan CO (2012) A polynomial approximation for arbitrary functions. *Applied Mathematics Letters* 25(11):1947–1952. DOI:10.1016/j.aml.2012.03.007
- David M (1988) *Handbook of applied advanced geostatistical ore reserve estimation*. Elsevier, Amsterdam
- De Iaco S, Maggio S (2011) Validation techniques for geological patterns simulations based on variogram and multiple-point statistics. *Math Geosci* 43(4):483–500. DOI:10.1007/s11004-011-9326-9
- Deutsch CV, Journel AG (1992) *GSLIB geostatistical software library and user's guide*. 2.0, 2nd edn. Oxford University Press, New York
- Dimitrakopoulos R, Luo X (2004) Generalized sequential gaussian simulation on group size v and screen-effect approximations for large field simulations. *Math Geol* 36(5):567–591. DOI:10.1023/B:MATG.0000037737.11615.df
- Dimitrakopoulos R, Mustapha H, Gloaguen E (2010) High-order statistics of spatial random fields: Exploring spatial cumulants for modeling complex non-Gaussian and non-linear phenomena. *Math Geosci* 42(1):65–99. DOI:10.1007/s11004-009-9258-9
- Goodfellow R, Albor Consuegra F, Dimitrakopoulos R, Lloyd T (2012) Quantifying multi-element and volumetric uncertainty, Coleman McCreeley deposit, Ontario, Canada. *Computers & Geosciences* 42(0):71–78. DOI:10.1016/j.cageo.2012.02.018

- Goovaerts P (1997) Geostatistics for natural resources evaluation. Applied geostatistics series. Oxford University Press, New York; Oxford, England
- Guardiano F, Srivastava RM (1993) Multivariate geostatistics: Beyond bivariate moments. In: Soares A (ed) Geostatistics Tróia '92, vol 5. Quantitative Geology and Geostatistics. Springer Netherlands, pp 133–144. DOI:10.1007/978-94-011-1739-5_12
- Honarkhah M, Caers J (2010) Stochastic simulation of patterns using distance-based pattern modeling. *Math Geosci* 42(5):487–517. DOI:10.1007/s11004-010-9276-7
- Huang T, Lu D-T, Li X, Wang L (2013) GPU-based SNESIM implementation for multiple-point statistical simulation. *Computers & Geosciences* 54:75–87. DOI:10.1016/j.cageo.2012.11.022
- Johnson ME (1987) Multivariate generation techniques. In: Multivariate Statistical Simulation. John Wiley & Sons, Inc., pp 43–48. DOI:10.1002/9781118150740.ch3
- Journal A (1994) Modeling uncertainty: Some conceptual thoughts. In: Dimitrakopoulos R (ed) Geostatistics for the Next Century, vol 6. Quantitative Geology and Geostatistics. Springer Netherlands, pp 30–43. DOI:10.1007/978-94-011-0824-9_5
- Journal AG (1997) Deterministic geostatistics: A new visit. In: Baafi EY, Schofield NA (eds) Geostatistics Wollongong, vol 1. Kluwer Academic, Dordrecht, pp 292–301
- Journal AG (2003) Multiple-point geostatistics: A state of the art. Unpublished Stanford Center for Reservoir Forecasting paper
- Li X, Mariethoz G, Lu D, Linde N (2016) Patch-based iterative conditional geostatistical simulation using graph cuts. *Water Resources Research* 52(8):6297–6320. DOI:10.1002/2015WR018378
- Mao S, Journal A (1999) Generation of a reference petrophysical/seismic data set: the Stanford V reservoir. 12th Annual Report, Stanford Center for Reservoir Forecasting. Stanford, USA
- Mariethoz G, Renard P, Straubhaar J (2010) The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research* 46(11):W11536. DOI:10.1029/2008WR007621
- Minniakhmetov I, Dimitrakopoulos R (2017a) A high-order, data-driven framework for joint simulation of categorical variables. In: Gómez-Hernández JJ, Rodrigo-Ilarri J, Rodrigo-Clavero ME, Cassiraga E, Vargas-Guzmán JA (eds) Geostatistics Valencia 2016. Springer International Publishing, Cham, pp 287–301. DOI:10.1007/978-3-319-46819-8_19
- Minniakhmetov I, Dimitrakopoulos R (2017b) Joint high-order simulation of spatially correlated variables using high-order spatial statistics. *Math Geosci* 49(1):39–66. DOI:10.1007/s11004-016-9662-x
- Minniakhmetov I, Dimitrakopoulos R (2018) High-Order spatial simulation using Legendre-Like orthogonal splines. *Mathematical Geosciences* (Minor revision)
- Mustapha H, Dimitrakopoulos R (2010a) Generalized Laguerre expansions of multivariate probability densities with moments. *Computers & Mathematics with Applications* 60(7):2178–2189. DOI:10.1016/j.camwa.2010.08.008
- Mustapha H, Dimitrakopoulos R (2010b) High-order stochastic simulation of complex spatially distributed natural phenomena. *Math Geosci* 42(5):457–485. DOI:10.1007/s11004-010-9291-8
- Mustapha H, Dimitrakopoulos R (2010c) A new approach for geological pattern recognition using high-order spatial cumulants. *Computers & Geosciences* 36(3):313–334. DOI:10.1016/j.cageo.2009.04.015
- Mustapha H, Dimitrakopoulos R (2011) HOSIM: A high-order stochastic simulation algorithm for generating three-dimensional complex geological patterns. *Computers & Geosciences* 37(9):1242–1253. DOI:10.1016/j.cageo.2010.09.007
- Osterholt V, Dimitrakopoulos R (2007) Simulation of wireframes and geometric features with multiple-point techniques: Application at Yandi iron ore deposit, Australia. In: Orebody modelling and strategic mine planning, vol 14. 2 edn. AusIMM Spectrum Series, pp 51–60
- Parra Á, Ortiz J (2011) Adapting a texture synthesis algorithm for conditional multiple point geostatistical simulation. *Stoch Environ Res Risk Assess* 25(8):1101–1111. DOI:10.1007/s00477-011-0489-1
- Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGEMS : a user's guide. Cambridge University Press, Cambridge, UK; New York

- Rosenblatt M (1952) Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* 23(3):470–472. DOI:10.2307/2236692
- Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. *Math Geol* 34(1):1–21. DOI:10.1023/A:1014009426274
- Strebelle S, Cavelius C (2014) Solving speed and memory issues in multiple-point statistics simulation program SNESIM. *Math Geosci* 46(2):171–186. DOI:10.1007/s11004-013-9489-7
- Wang H, Xiang S (2012) On the convergence rates of Legendre approximation, vol 81. DOI:10.2307/23267976
- Wu J, Boucher A, Zhang T (2008) A SGeMS code for pattern simulation of continuous and categorical variables: FILTERSIM. *Computers & Geosciences* 34(12):1863–1876. DOI:10.1016/j.cageo.2007.08.008
- Xu W (1996) Conditional curvilinear stochastic simulation using pixel-based algorithms. *Math Geol* 28(7):937–949. DOI:10.1007/BF02066010
- Zarowski CJ (2004) Orthogonal polynomials. In: *An Introduction to numerical analysis for electrical and computer engineers*. John Wiley & Sons, Inc., pp 207–250. DOI:10.1002/0471650412.ch5
- Zhang T, Switzer P, Journel A (2006) Filter-based classification of training image patterns for spatial simulation. *Math Geol* 38(1):63–80. DOI:10.1007/s11004-005-9004-x