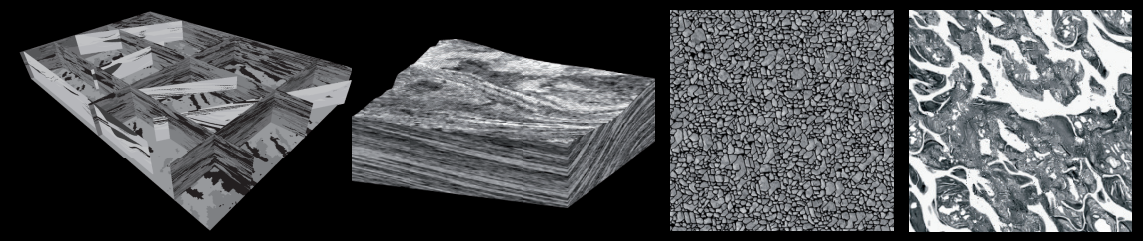
Geological stochastic imaging for aquifer characterization

G. Mariethoz

=



Geological stochastic imaging
for aquifer characterization

Grégoire Mariethoz

Université de Neuchâtel unine

Université de Neuchâtel unine

FACULTE DES SCIENCES
Secrétariat-Décanat de la faculté
- Rue Emile-Argand 11
- CP 158
- CH-2009 Neuchâtel

# IMPRIMATUR POUR LA THESE

## Geological stochastic imaging for aquifer characterization

**Grégoire Mariethoz**

---

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel,
sur le rapport des membres du jury

MM. P. Renard (directeur de thèse), O. Besson, P. Atkinson (Southampton UK),
J. Caers (Stanford University, USA), A. Boucher (Stanford University USA)
et Lin Y. Hu (Houston USA)

autorise l'impression de la présente thèse.

Neuchâtel, le 20 août 2009

Le doyen :
F. Kessler

UNIVERSITE DE NEUCHATEL
FACULTE DES SCIENCES
Secrétariat - décanat de la faculté
Rue Emile-Argand 11 - CP 158
CH-2009 Neuchâtel

■ Téléphone : +41 32 718 21 00 ■ E-mail : secretariat.sciences@unine.ch ■ www.unine.ch/sciences

This PhD was funded by the Swiss National Science Foundation (SNSF grant PP002-1065557) and the Swiss Confederation's Innovation Promotion Agency (CTI project 8836.1 PFES-ES).

*A mes filles, Nina et Anahi*

# Remerciements

Je tiens remercier sincèrement à mon directeur de thèse, Philippe Renard, qui m'a procuré un encadrement et une disponibilité exceptionnels durant les trois années qu'à duré mon travail de doctorat. Je n'ai jamais autant appris, et il en porte le mérite. Nos discussions fréquentes et informelles ont été le berceau de nombreuses idées développées dans ce travail. En me permettant d'assister à de multiples conférences internationales et à suivre de nombreux cours, il a élargi mes horizons au-delà de mes espérances et largement accompli sa tâche d'enseignant.

Même s'il n'est plus là pour les recevoir, j'adresse mes remerciements à Martin Burkhard qui le premier m'a intéressé à la géologie, m'a permis d'orienter mes études dans cette voie, qui m'a prodigué de nombreux conseils en plus de ses cours passionnants, et sans qui je ne serais pas hydrogéologue.

Je remercie également Julien Straubhaar pour les innombrables heures où nous avons brassé des arguties algorithmiques, au risque de passer pour une paire d'hirsutes. Je lui suis reconnaissant de m'avoir initié à la programmation parallèle, et aussi pour toutes les fois où, avec une patience infinie, il m'a sorti de problèmes mathématiques ou informatiques. Sans lui, ce travail n'aurait pas vu le jour sous cette forme.

Je tiens également à remercier Olivier Besson, Roland Froidevaux, Alexandre Walgenwitz, Denis Allard, Pierre Biver et Tatiana Chugunova qui m'ont énormément apporté dans le cadre du projet CTI, sans lequel je n'aurai jamais commencé à travaillé sur les méthodes multi-points.

Merci à Alexandre Boucher pour m'avoir prodigué de précieux conseils et commentaires qui ont largement influencé ma dernière année de thèse. Merci aussi à Rami Younis qui a su voir le potentiel des processeurs graphiques.

Je suis particulièrement reconnaissant envers Olivier Jaquet, Albert Tarantola, John Doherty, David Ginsbourger, Jacques Rivoirard, Didier Renard et Hélène Beucher dont les enseignements reçus lors de divers cours ont largement influencé mon travail. De prolifiques discussions avec Colin Daly, Jef Caers, Lin Hu, Peter Atkinson, Céline Scheidt, Tapan Mukerji, Antoine Bertoncello, Sebastien Strebelle, David Ardia et Jesus Carrera ont également eu un grand impact sur mon travail. Je tiens à remercier particulièrement Jef Caers de me permettre de poursuivre mon travail sous la forme d'un postdoc à Stanford.

Merci à Peter Bayer et à Philipp Blum pour leur invitation et pour les débats scientifiques autour des bières de Tuebingen. Un grand merci à Baptiste Dafflon, Jaouher Kerrou, Damian Glenz, Andrea Borghi, Vincent Badoux, Alessandro Communian, Andrés Alcolea, Alain Pochon, Jean-Marie Lepioufle, Mehrdad Honarkhah, Diana dell'Arciprete et Mike Ronayne qui, en me faisant découvrir leurs propres recherches, m'ont également beaucoup appris.

Je remercie Fabien Cornaton pour la patience dont il a fait preuve lorsque j'avais des difficultés à utiliser Groundwater, et pour m'avoir sauvé des murènes à Sur. Merci à Christian Ghasarian pour m'avoir dissuadé de faire une thèse en ethnologie et à Charles Elwood Shannon pour ce qui était à son sens une petite astuce pratique. Je remercie Mathias Meylan pour avoir réveillé chez moi un vieil intérêt pour l'informatique, ce qui m'a bien aidé lors de cette thèse, ainsi que Corentin Zumwald pour sa vieille amitié. Merci à Mathieu Beck et Nicolas Coppo pour une inoubliable et dangereuse descente du Soliat et pour m'avoir fait participer à une étude géophysique désintéressée. Merci à Robin Dufour pour son show permanent et pour un mandat au Pérou. Merci également à Tristan Ibrahim pour avoir souvent pallié aux défaillances bibliographiques de l'Université de Neuchâtel.

Je tiens à remercier tout particulièrement Romain Sonney pour l'achat d'un jeu de backgammon et François Zwahlen pour le financement d'un billard, deux acquisitions qui ont égayé mes années de thèse au CHYN. Merci à Nicolas Bouvier, Hugo Pratt et Terry Gilliam pour leur contribution. Merci à ma mère, à ma sœur Céline et également à Daniel pour leurs encouragements. Et finalement, merci à Mirjam pour m'avoir soutenu, supporté et pour avoir partagé avec moi ces riches années.

Ces remerciements ne sont sans doute pas exhaustifs, je remercie donc encore tous ceux que j'ai oublié.

# Abstract

Accurately modeling connectivity of geological structures is critical for flow and transport problems. Using multiple-points simulations is one of the most advanced tools to produce realistic reservoir structures. It proceeds by considering data events (spatial arrangements of values) derived from a training image (TI). The usual method consists in storing all the data events of the TI in a database, which is used to compute conditional probabilities for the simulation. Instead, the Direct Sampling method (DS) proposed in this thesis consists in sampling directly the TI for a given data event. As soon as the data event in the TI matches the data event at the node to simulate, the value at its central node is directly pasted in the simulation. Because it accommodates data events of varying geometry, multi-grids are not needed. The method can deal with categorical and continuous variables and can be extended to multivariate cases. Therefore, it can handle new classes of problems.

Different adaptations of the DS are proposed. The first one is aimed at reconstructing partially informed images or datasets. Instead of inferring data events from a TI, a training dataset is used. If the density of measurements is high enough, significant non-parametric spatial statistics can be derived from the data, and the patterns found in those data are mimicked without model inference. Therefore, minimum assumptions are made on the spatial structure of the reconstructed fields. Moreover, very limited parameterization is needed. The method gives good results for the reconstruction of complex 3D geometries from relatively small datasets.

Another adaptation of the DS algorithm is aimed at performing super-resolution of coarse images. DS is used to stochastically simulate the structures at scales smaller than the measurement resolution. These structures are inferred using a hypothesis of scale-invariance on the spatial patterns found at the coarse scale. The approach is illustrated with examples of satellite imaging and digital photography.

Parallelization is another important topic treated in this thesis. The size of simulation grids used for numerical models has increased by many orders of magnitude in the past years. Efficient pixel-based geostatistical simulation algorithms exist, but for very large grids and complex spatial models, computational burden remains heavy. As cluster computers become widely available, using parallel strategies is a natural step for increasing the usable grid size and the complexity of the models. These strategies must take profit of the possibilities offered by machines with a large number of processors. On such machines, the bottleneck is often the communication time between processors. This thesis presents a strategy distributing grid nodes among all available processors while minimizing communication and latency times. It consists in centralizing the simulation on a master processor that calls other slave processors as if they were functions simulating one node every time. The key is to decouple the sending and the receiving operations to avoid synchronization. Centralization allows having a conflict management system ensuring that nodes being simulated simultaneously do not interfere in terms of neighborhood. The strategy is computationally efficient and is versatile enough to be applicable to all random path based simulation methods.

In addition to the preceding topics, a new cosimulation algorithm is proposed for simulating a primary attribute using one or several secondary attributes known exhaustively on the domain. This problem is frequently encountered in surface and groundwater hydrology when a variable of interest is measured only at a discrete number of locations and when a secondary variable is mapped by indirect techniques such as geophysics or remote sensing. In the proposed approach, the correlation between the two variables is modeled by a joint probability distribution function. A technique to construct such relations using latent variables and physical laws is proposed when field data are insufficient. The simulation algorithm proceeds sequentially. At each node of the grid, two conditional probability distribution functions (cpdf) are inferred. The first is inferred in a classical way from the neighboring data of the main attribute and a model of its spatial variability. The second is inferred directly from the joint probability distribution function of the two attributes and the value of the secondary attribute at the location to be simulated. The two distribution functions are combined by probability aggregation to obtain the local cpdf from which a value is randomly drawn. Various examples using synthetic and remote sensing data demonstrate that the method is more accurate than the classical collocated cosimulation technique when a complex relation links the two attributes.

# Table of contents

## Chapter 3

Reconstruction of incomplete data sets or images using Direct Sampling



## Chapter 4

Super-resolution using multiple-points statistics

## Chapter 5

A general parallelization strategy for random path based geostatistical simulation methods



## Chapter 6

Integrating collocated auxiliary parameters in geostatistical simulations using joint probability distributions and probability aggregation

## Chapter 7

Concluding comments



## Appendixes

## A

Truncated plurigaussian simulations to characterize aquifer heterogeneity

## B

Reducing the impact of a desalination plant using stochastic modeling and optimization techniques



## C

Improving the performance of the Direct Sampling algorithm

**D**

A Matlab® code for Direct Sampling

# Chapter 1

# Introduction

*All medals carry two faces, no convenience come free of limitations. In a time of massive computing power, there is no more justification to accept blindly the limitations associated to parameter-poor Gaussian-related models. Data and prior geological concepts, no matter how complex they are, should be let to speak for themselves.*

*Journel and Zhang (2006)*

## 1. Motivations

Aquifer characterization calls for Stochastic methods because of their ability to describe the ubiquitous geological uncertainty (De Marsily, *et al.*, 2005). This added knowledge is crucial for hydrogeological problems. Indeed, the connectedness and the disposition of the underground structures are the main factors driving flow and especially transport processes (Gómez-Hernández and Wen, 1998). Stochastic methods have also proven to be invaluable tools for management purposes. They provide decision makers with quantitative descriptions of the consequences of geological uncertainty for specific problems, such as designing a remediation strategy for contaminated sites, planning long-term water resources or evaluating the effect of human activity on ecosystems.

The field of geostatistics was initiated with Earth Sciences applications (Matheron, 1965; Journel, 1974). The use of geostatistical methods rapidly spread and they are now used in disciplines such as meteorology, biology, epidemiology or fishery (e.g. Goovaerts, 1997). Although the fields of applications have broadened, Earth Science have continued developing ever more diversified and sophisticated techniques for the description of facies and flow-related properties such as hydraulic conductivity or porosity (e.g. Le Loc'h and Galli, 1994; Koltermann and Gorelick, 1996; Carle and Fogg, 1997; e.g. Armstrong, *et al.*, 2003). The most recent advances in geostatistics are the training-image based methods that allow including qualitative geological knowledge in the models (Guardiano and Srivastava, 1993; Strebelle, 2002; Zhang, *et al.*, 2006; Arpat and Caers, 2007; Hu and Chugunova, 2008; Straubhaar, *et al.*, 2008). The training image can be seen as a geologically-driven prior model; it can be multiGaussian, but mutliGaussianity is not a necessity (Journel and Zhang, 2006).

Although these methods are new, they are widely applied in the petroleum industry (Strebelle, 2008), and application cases are starting to appear in hydrogeology (e.g. Feyen and Caers, 2006; e.g. Ronayne and Gorelick, 2006). Compared to the traditional parametric methods that were getting more and more sophisticated and complex, training image based techniques can be seen as a stunning return to simplicity (at least conceptually, not in terms of model or algorithm simplicity). The Direct Sampling method presented in this thesis tends toward even more simplicity for the algorithms used, their functional concepts and their parameterization.

In addition to simplicity, the method proposes a different way of manipulating probabilities than is commonly used in geostatistics. It does not infer probability distributions, but instead uses the concept of distance between data events. Distances offer a high degree of flexibility and allow dealing with both categorical and continuous variables. This flexibility should allow using these methods in domains way beyond hydrogeology. Examples are provided for the generation of synthetic rainfall and the processing of digital photography, but several other applications can be envisioned, such as the simulation of financial indices, the probabilistic mapping of archeological remains, the restoration of damaged images (or movies), as well as meteorological short-term prediction.

Renard (2007) shows that the use of stochastic methods in hydrogeology is under-estimated by practitioners, and provides insights on the causes of this disinterest. One major reason mentioned is the complexity of most geostatistical approaches, which makes it time-consuming (and therefore costly) to get familiar with and to apply. The simplicity of the approaches proposed in this thesis is a step towards improving this situation. For example, model inference is reduced to finding an appropriate training image that can be either categorical or continuous, uni- or multivariate, stationary or non-stationary. These inherent characteristics of the training image are reproduced on the domain to simulate without further adjustment of parameters. If large datasets are available, the simulation becomes a reconstruction problem (see chapter 3) and the training image is not necessary any more. The use of stochastic methods is then made extremely straightforward. Such ease of use could be a major factor to draw the attention of practitioners towards stochastic hydrogeology.

Another reason mentioned by Renard (2007) for the lack of investment in stochastic hydrogeology is the insufficient realism of the generated models, from a geological point of view. For many geologists, the standard multi-Gaussian model does not adequately represent geological variability. Training-image based methods are clearly more oriented towards integrating geological concepts. By allowing the use of training images with continuous properties, Direct Sampling is another tool for translating geological concepts into geostatistical models.

Building geological models often involves to integrate several types of information coming from various sources. Additional information can come from geophysical measurements, remote sensing images, expert's opinion, etc. The relationship between this auxiliary data and the parameters governing the model output (typically hydraulic conductivity, storativity, porosity, etc) is often not straightforward. Linear correlations have shown their limits. Natural processes are

seldom linearly related, and the failure to account for non-linearity gives unrealistic results, yielding poor predictions. Direct Sampling allows accounting for complex correlations between variables without linear assumptions. In addition, a method for integrating any known relationship between variables is presented in this thesis. It is based on a probability aggregation approach (Bordley, 1982) stemming from management science.

The calculation time devoted to the generation of stochastic images has not been an issue until recently because it was negligible compared to the time needed for solving a flow and transport problem. The advent of CPU-demanding geostatistical algorithms and the use of increasingly large grids have now made this question central. With multiple-points methods, the time needed for obtaining a realistic parameters field is comparable to the time taken by a steady-state flow simulation.

When addressing inverse problems and for Monte-Carlo analysis, large numbers of realizations have to be produced, and the question of computation time is crucial. Actually, the existence of a limited computation capability is the only reason why the Metropolis-Hastings algorithm is not systematically used for all hydrogeological inverse problems (Tarantola, 2005). Finding less CPU-demanding approximations for determining posterior distributions is a very broad and active research topic (Carrera and Neuman, 1986; De Marsily*, et al.*, 1999; Doherty, 2003).

Computing cost is also one of the reasons mentioned by Renard (2007) to explain why stochastic methods are not more often used in practice. CPU-time is the bottleneck of the aquifer characterization workflow, and acceleration strategies are needed. Parallel capabilities are available on virtually all modern computers, and taking advantage of this feature allows significant gains in computation time. Parallel solvers for flow and transport problems already exist and are widely distributed. Developing parallel stochastic imaging algorithms is a natural research direction for the field of geostatistics. One such method is proposed in this thesis. It is flexible enough to be used on any kind of parallel machine. Other ways of parallelizing geostatistical algorithms may be envisioned and new methods are expected to appear in the near future.

## 2. Structure of the thesis

This thesis presents new methods for generating stochastic images. These methods allow to:

- Produce images by borrowing and recomposing patterns coming either from a training image, either from a data set, either from a fractal extrapolation of a lower resolution image. The core idea is simple and old (Shannon, 1948). Such techniques are developed in chapters 2,3 and 4.
- Reduce the time needed to produce images through parallelization. Chapter 5 is devoted to this aspect.
- Generate parameters fields presenting any kind of pair-wise correlation with another given parameter field. This problem is treated in Chapter 6.

Appendixes A and B are already published papers related to applications of stochastic methods in hydrogeology. Both are typical applications of stochastic methods to real-case hydrogeological problems. The geostatistical methods used are traditional 2-point statistics, but such problems could very well be addressed with the Direct Sampling method. These works were carried on during the time of the PhD.

Appendix A illustrates a workflow for the stochastic characterization of contamination associated to a leaking waste deposit. Geostatistical simulations are used (truncated plurigaussian in this case) because of high geological uncertainty and extreme heterogeneity.

Appendix B is the result of a fruitful collaboration with industrial partners. It combines inverse stochastic characterization of an aquifer with a non-linear optimization problem for positioning wells at the most productive locations.

Further discussion on the Direct Sampling (chapter 2) is provided in the last two appendixes. Appendix C describes additions to the algorithm that increase CPU efficiency. Appendix D presents a small Matlab® code that illustrates the core of the method.

## 3. References


Armstrong, M., Galli, A. G., Loc'h, G. L., Geoffroy, F., and Eschard, R. (2003), *Plurigaussian Simulations in Geosciences*, Springer.

Arpat, B., and Caers, J. (2007), *Conditional Simulations with Patterns*, Mathematical Geology, 39, 2, 177-203.

Bordley, R. E. (1982), *A multiplicative formula for aggregating probability assessments*, Management Science, 28, 10, 1137-1148.

Carle, S. F., and Fogg, G. E. (1997), *Modeling spatial variability with one and multi-dimensional continuous Markov chains*, Mathematical Geology, 7, 29, 891-918.

Carrera, J., and Neuman, S. (1986), *Estimation of aquifer parameters under transient and steady-state conditions, 2. Uniqueness, stability and solution algorithms.*, Water Resour. Res., 22, 2, 211-227.

De Marsily, G., Delay, F., Gonçalvès, J., Renard, P., Teles, V., and Violette, S. (2005), *Dealing with spatial heterogenity*, Hydrogeology Journal, 13, 1, 161-183.

De Marsily, G., Delhomme, J.-P., Delay, F., and A., B. (1999), *Regards sur 40 ans de problèmes en hydrogéologie*, Acad. Sci. Paris, Sciences de la terre et des planètes, 329, 73-87.

Doherty, J. (2003), *Ground water model calibration using pilot points and regularization*, Ground Water, 41, 2, 170-177.

Feyen, L., and Caers, J. (2006), *Quantifying geological uncertainty for flow and transport modelling in multi-modal heterogenesous formations*, Advances in Water Resources 29, 6, 912-929.

Gómez-Hernández, J. J., and Wen, X.-H. (1998), *To be or not to be multi-gaussian? A reflection on stochastic hydrogeology*, Advances in Water Resources, 21, 1, 47-61.

Goovaerts, P. (1997), *Geostatistics for Natural Resources evaluation*, Oxford University Press, Oxford.

Guardiano, F., and Srivastava, M. (1993), *Multivariate geostatistics: Beyond bivariate moments*, in *Geostatistics-Troia*, pp. 133-144, Kluwier Academic.

Hu, L., and Chugunova, T. (2008), *Multiple-Point Geostatistics for Modeling Subsurface Heterogeneity: a Comprehensive Review*, Water Resour. Res., 44, W11413.

Journel, A. (1974), *Geostatistics for conditional simulation of ore bodies*, Econ. Geol., 69, 5, 673-687.

Journel, A., and Zhang, T. (2006), *The Necessity of a Multiple-Point Prior Model*, Mathematical Geology, 38, 5, 591-610.

Koltermann, C., and Gorelick, S. (1996), *Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches*, Water Resour. Res., 32, 9, 2617-2658.

Le Loc'h, G., and Galli, A. G. (1994), *Improvement in the truncated Gaussian method: combining several Gaussian functions*, paper presented at Ecmor 4, 4th European Conference on the Mathematics of Oil Recovery, Roros, Norway.

Matheron, G. (1965), *Les variables régionalisées et leur estimation*, Masson.

Renard, P. (2007), *Stochastic hydrogeology: what professionals really need?*, Ground Water, 45, 5, 531-541.

Ronayne, M., and Gorelick, S. (2006), *Effective permeability of porous media containing branching channel networks*, Physical Review E, 73, 026305, 1-10.

Shannon, C. E. (1948), *A mathematical theory of communication*, The Bell system technical journal, 27, 379-423.

Straubhaar, J., Walgenwitz, A., Renard, P., and Froidevaux, R. (2008), *Optimization issues in 3D multipoint statistics simulation*, paper presented at Geostats 2008, 1-5 Dec. 2008, Santiago, Chile.

Strebelle, S. (2002), *Conditional simulation of complex geological structures using multiple point statistics.*, Mathematical Geology, 34, 1, 1-22.

Strebelle, S. (2008), *Multiple-Point Geostatistics: from Theory to Practice*, paper presented at 21st SCRF Meeting, Stanford University, May 8-9, 2008.

Tarantola, A. (2005), *Inverse Problem Theory and Methods for Parameter estimation*, Society for Industrial and Applied Mathematics.

Zhang, T., Switzer, P., and Journel, A. (2006), *Filter-Based Classification of Training Image Patterns for Spatial Simulation*, Mathematical Geology, 38, 1, 63-80.

# Chapter 2

# The Direct Sampling method to perform multiple-points geostatistical simulations*

*"The […] samples were constructed by the use of a book of random numbers in conjunction with […] a table of letter frequencies. This method might have been continued […], since digram, trigram and word frequency tables are available, but a simpler equivalent method was used. […] one opens a book at random and selects a letter at random on the page. This letter is recorded. The book is then opened to another page and one reads until this letter is encountered. The succeeding letter is then recorded. Turning to another page this second letter is searched for and the succeeding letter recorded, etc. […]. It would be interesting if further approximations could be constructed, but the labor involved becomes enormous at the next stage."*

*Claude Elwood Shannon, A Mathematical Theory of Communication, 1948*

---

* This chapter has been submitted for publication in Water Resources Research as:
Mariethoz, G., P. Renard, J. Straubhaar. The Direct Sampling method to perform multiple-points geostatistical simulations.
The algorithms described are protected by the international patent 2008WO-EP009819.

**Abstract** Multiple-point geostatistics is one of the most general statistical frameworks to model spatial fields displaying a wide range of complex structures. In particular, it allows controlling connectivity patterns which have a critical importance for groundwater flow and transport problems. This approach proceeds by considering data events (spatial arrangements of values) derived from a training image (TI). All data events found in the TI are usually stored in a catalogue, which is scanned to compute conditional probabilities for the simulation. Instead, we propose to sample directly the training image for a given data event. As soon as the data event in the training image matches the data event in the neighborhood of the node to simulate, the value at its central node is assigned in the simulation. This technique is statistically equivalent to previous implementations but, because it relaxes the need to compute explicitly the conditional probabilities, it allows to extend the application of the standard theory from categorical to continuous and multivariate variables. This method can be used for the simulation of geological heterogeneity, accounting or not for indirect observations such as geophysics, but it can also be used to simulate any variable that has non multi-Gaussian features which may have an important impact on hydrological processes. Computationally, it is fast, easy to parallelize, and parsimonious in memory needs. Because data events of varying geometry can be accommodated, multi-grids are not needed.

## 1. Introduction

Geological heterogeneity has a critical influence on flow and transport behavior in aquifer formations. This heterogeneity, complex because produced by natural processes, is often characterized using geostatistics. Hydrogeology has always been a pioneering discipline for the application of stochastic methods. As soon as the sixties, early geostatistical developments were specifically aimed at investigating the impact of heterogeneity on flow and transport through porous media (Matheron, 1966; , 1967). These developments were quickly extended in the arena of hydrogeology (Freeze, 1975; e.g. Dagan, 1976; Gelhar, et al., 1977; Delhomme, 1979), where stochastic methods became a generalized (though debated) research topic (Christakos, 2004; Renard, 2007). Since these early works, research in hydrogeology and geostatistics stayed closely related. Numerous techniques now exist to generate alternative images of the underground structures, which serve as input to groundwater flow and transport simulators. Several review papers summarize the accomplishments realized in the now well-established domain of stochastic hydrogeology (Neuman, 1984; Dagan, 1986; Gelhar, 1986; Koltermann and Gorelick, 1996; Carrera, et al., 2005; De Marsily, et al., 2005; Hu and Chugunova, 2008). Stochastic hydrogeology deals with a wide range of problems.

For example, Mariethoz, et al. (2009) used truncated plurigaussian simulations to assess the contamination risk related to a waste disposal. Another application related to water supply is the characterization of an aquifer with inverse methods and the use of optimization techniques to define an optimum pumping scheme for a desalinization plant (Alcolea, et al., 2009). Numerous other examples can be found in the literature, such as Ronayne, et al. (2008) who use inverse stochastic modeling to identify specific geologic features influencing the results of pumping tests.

In recent years, traditional approaches using parametric statistics have been questioned, because such statistical models fail to represent some important features controlling flow and transport behavior. For example, Zinn and Harvey (2003) and Sánchez-Vila, et al. (1996) have shown very clearly how the multi-Gaussian model of heterogeneity leads to a very specific type of upscaled behavior that may be right or wrong depending on the type of connectivity patterns displayed by the hydraulic conductivity field. Kerrou et al. (2008) have also shown that the choice of a multi-Gaussian model coupled with inverse techniques may lead to inaccurate and even biased forecasts when the reality is not multi-Gaussian.

More generally, the issue of the connectivity of multi-Gaussian fields and its consequences on flow and transport problems has been discussed at length in the framework of the characterization of heterogeneity (Journel and Alabert, 1990; Journel and Deutsch, 1993; Koltermann and Gorelick, 1996; Gómez-Hernández and Wen, 1998; Western, et al., 2001).

Addressing non-multi-Gaussian features has motivated the progressive development of a broad range of alternative methods (De Marsily, et al., 2005). Among those techniques, multiple-points statistics (Guardiano and Srivastava, 1993) has prevailed as one of the most promising. A detailed review of the approach has been done by Hu and Chugunova (2008). One of the most cited implementation of multiple-points simulations is the *snesim* algorithm (Strebelle, 2002) that is now increasingly used in the oil industry (Caers, et al., 2003; Strebelle, et al., 2003; Liu, et al., 2004; Aitokhuehi and Durlofsky, 2005; Hoffman and Caers, 2007) and in hydrogeology (Feyen and Caers, 2006; Renard, 2007; Chugunova and Hu, 2008; Huysmans and Dassargues, 2008; Ronayne, et al., 2008).

Although the method is gaining popularity, it still presents several shortcomings. Some of the most acute ones are the impossibility to simulate continuous variables and to perform cosimulations. The prohibitive memory requirements of the algorithm make it difficult to apply in 3D cases. Moreover, the method is very CPU demanding and difficult to parallelize.

In this paper, we propose an alternative multiple-points simulation algorithm (the Direct Sampling) that palliates these issues and offers a potential for new fields of application. It can deal with both categorical and continuous variables, can address non-stationarity in several ways, does not require large amounts of memory and is easy to parallelize. The algorithm is simple, lightweight and easy to implement.

The first part of the paper provides a technical overview of multiple-points geostatistics and highlights the novel aspects of the Direct Sampling method (DS). The second section is a detailed description of the DS algorithm. The following sections illustrate with examples the possibilities offered by the method, such as simulating continuous properties, addressing multivariate problems and dealing with non-stationarity.

To obtain a good compromise between numerical efficiency and quality of pattern reproduction, we also propose a recursive syn-processing method inspired from existing post-processing algorithms (Strebelle and Remy, 2005; Stien, et al., 2007; Suzuki and Strebelle, 2007). The syn-processing is applied in real time and aims at ensuring a maximum of consistency between the patterns found in the training image and those simulated by the algorithm regardless of their size, shape and complexity. Syn-processing is applied in conjunction with DS, but could be used with any other multiple-points simulation algorithm.

## 2. Background on multiple-points geostatistics

Multiple-points geostatistics is based on two radical conceptual changes that were first formalized by Guardiano and Srivastava (1993). The first one is to state that data sets are usually not sufficient to infer all the statistical features that control what the modeler is interested in. For example, based only on point data, it is impossible to know whether the high values of hydraulic conductivity are connected or belong to isolated blocks (Gómez-Hernández and Wen, 1998). Therefore any statistical inference based only on the analysis of point data (even if it uses complex statistics) will just be blind to that characteristic of the underlying field. In terms of flow behavior, the consequences of such errors can be drastic. The only way to build a statistical model of the heterogeneity that displays a certain type of connectivity is to add external information not contained in the data itself. One approach is to devise a model that contains in its structure a certain type of connectivity as it is possible with Boolean techniques. The proposal of Guardiano and Srivastava (1993) is to use a training image (TI), i.e. a grid containing spatial patterns deemed representative of the spatial structures to simulate. The training image can be viewed

as a conceptual model of the heterogeneity in the case of reservoir characterization but should be seen more generally as an explicit prior model (Journel and Zhang, 2006). The statistical model is then based not on the data but on the training image. The choice of the training image allows the modeler to integrate external information about spatial variability such as geological knowledge.

The second radical change is to adopt a non-parametric statistical framework allowing to deal with multiple-points statistics instead of two point-statistics (Guardiano and Srivastava, 1993). The multiple-point statistics are expressed as the cumulative density functions for the random variable $Z(\mathbf{x})$ conditioned to local data events $\mathbf{d}_n=\{Z(\mathbf{x}_1), Z(\mathbf{x}_2),\cdots, Z(\mathbf{x}_n)\}$, i.e. the values of $Z$ in the neighboring nodes $\mathbf{x}_i$ of $\mathbf{x}$:

$$F(z,\mathbf{x},\mathbf{d}_n) = \text{Prob}\{Z(\mathbf{x}) \le z \mid \mathbf{d}_n\}. \quad (1)$$

The simulation proceeds sequentially. At each successive location, the conditional cumulative distribution function (ccdf) $F(z,\mathbf{x},\mathbf{d}_n)$ is conditioned to both the previously simulated nodes and the actual data. A value for $Z(\mathbf{x})$ is drawn from the distribution and the algorithm proceeds to the next location. Because $F(z,\mathbf{x},\mathbf{d}_n)$ depends on the respective values and relative positions of all the neighbors of $\mathbf{x}$ simultaneously, it contains much more information than the usual two point statistics defined by a covariance function or variogram. To estimate the non parametric ccdf (1) at each location, Guardiano and Srivasta (1993) proposed to scan the training image for each new location. The method was inefficient and therefore could not be used in practice. A very elegant solution to that problem was then developed by Strebelle (2002). The technique consists in scanning the training image for all pixel configurations of a certain size (the template size) and to store their statistics in a catalogue of data events having a tree structure. This structure is then used to compute rapidly the conditional probabilities at each simulated node. In general, to limit the size of the tree in memory, the template size has to remain small and does not allow capturing large-scale features such as channels. To palliate this problem, Strebelle (2002) introduces multi-grids to simulate the large-scale structures first, and later the small-scale features. Although multi-grids allow good reproduction of the structures at different scales, it generates problems related to the migration of conditioning data at each multi-grid level. Artifacts may appear, especially with large datasets that cannot be fully used on the coarsest multi-grids levels.

Because all configurations of pixel values that are found in the TI are stored in the search tree, the algorithm is rapidly limited by the memory usage. More

precisely, the size of the template, the number of categories (usually representing lithofacies) and the degree of entropy of the training image directly control the size of the search tree and therefore control the memory requirement for the algorithm. In practice, these parameters are limited by the available memory especially for large 3D grids. For example, with 4 lithofacies and a template made of 30 nodes, there can be up to $4^{30}$ possible data events, which by far exceeds the memory limit of any present-day computer (although in practice, the number of data events is limited by the size of the TI). This imposes limits on the number of facies and the template size, and hence complex structures described in the TI can often not be properly reproduced. Straubhaar et al. (2008) mitigated this problem by storing multiple-points statistics in lists instead of tree structures. In addition, to account for non-stationarity either in the training image or in the simulation, it is necessary to include additional variables that further increase the demand for memory storage (Chugunova and Hu, 2008).

Traditional approaches can only deal with categorical variables, (although continuous variables can be discretized into classes (Strebelle, 2007), but then the question of how to define the classes arises). The direct estimation of (1) for a continuous variables is problematic. Zhang et al. (2006) propose an alternative method in which the patterns are projected (through the use of filter scores) into a smaller dimensional space in which the statistical analysis can be carried out. The resulting *filtersim* algorithm does not simulate nodes one by one sequentially, but proceeds by pasting groups of pixels (patches) into the simulation grid. It uses the concept of similarity measure between groups of pixels, and can be applied both to continuous or categorical variables. Another alternative technique based on image analysis and pattern recognition techniques is the *simpat* algorithm (Arpat and Caers, 2007). This last alternative technique tends to consider the simulation of heterogeneous fields as a pattern reproduction problem and not as a probabilistic problem. It should therefore be viewed as belonging to another class of methods even if it also uses the concept of training image.

In this paper, we adopt the point of view that generating simulations satisfying the ccdf expressed in equation (1) does not involve explicitly computing this ccdf. We therefore suggest that the technical difficulties involved in the computation of the ccdf can be avoided. Instead of storing and counting the configurations found in the training image, it is more convenient to sample directly the training image in a random manner but conditional to the data event. Mathematically, this is equivalent to using the training image (TI) to compute the ccdf and then drawing a sample from it. The resulting Direct Sampling (DS) algorithm is inspired from Shannon (1948)

who produces Markovian sequences of random English by drawing letters from a book conditionally to previous occurrences.

In addition, we introduce a distance (mismatch) between the data event observed in the simulation and the one sampled from the TI. During the sampling process, if a pattern is found that matches exactly or if the distance is lower than a given threshold, the sampling process is stopped and the value at the central node of the data event in the TI is directly used in the simulation. Choosing an appropriate measure of distance allows dealing with either categorical or continuous variables, and to accommodate complex multivariate problems such as correlation between categorical and continuous variables. This flexibility should allow using DS for a wide range of problems in hydrology. One example is the characterization of the spatial distribution of porosity and hydraulic conductivity for modeling flow and transport. Complex connectivity patterns of high and low values can be accounted for, which is crucial for transport problems. Moreover, both variables (porosity and hydraulic conductivity) can be simulated jointly in the multivariate framework, considering their relationship in a non-parametric way. It can also address issues that are usually beyond the scope of geostatistical methods, such as predicting meteorological processes that involve a high number of variables presenting complex interactions, and where a history of measurements can provide a suitable training image.

## 3. The Direct Sampling algorithm

The aim of the Direct Sampling method is to simulate a random function $Z(\mathbf{x})$. The input data are a simulation grid (SG) whose nodes coordinates are denoted $\mathbf{x}$, a training image (TI) whose nodes coordinates are denoted $\mathbf{y}$, and if available a set of $N$ conditioning data $z(\mathbf{x}_i), i \in [1, \cdots, N]$ such as for example borehole observations. The principle of the simulation algorithm is illustrated in Figure 1 and Figure 2 and proceeds as follows.

1) Each conditioning data is assigned to the closest grid node in the SG. If several conditioning data should be assigned to the same grid node, we assign the closest one to the center of the grid node.
2) Define a path through the remaining nodes of the SG. The path is a vector containing all the indices of the grid nodes that will be simulated sequentially. Random (where nodes are visited in a random manner, e.g. Strebelle, 2002),

unilateral (where nodes are visited in a regular order starting along one side of the grid, e.g. Daly, 2004) or any other path can be used.

3) For each successive location **x** in the path:
   a) Find the neighbors of **x**. They consist of a maximum of the *n* closest grid nodes $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ that were already assigned or simulated in the SG. If no neighbor is found for **x** (e.g. for the first node of an unconditional simulation), randomly take a node **y** in the TI and assign its value $Z(\mathbf{y})$ to $Z(\mathbf{x})$ in the SG. The algorithm can then proceed to the next node in the path.
   b) Compute the lag vectors $\mathbf{L} = \{\mathbf{h}_1, \cdots, \mathbf{h}_n\} = \{\mathbf{x}_1 - \mathbf{x}, \cdots, \mathbf{x}_n - \mathbf{x}\}$ defining the neighborhood of **x,** $\mathbf{N}(\mathbf{x}, \mathbf{L}) = \{\mathbf{x} + \mathbf{h}_1, \cdots, \mathbf{x} + \mathbf{h}_n\}$. For example, in Figure 1a the neighborhood of the grey pixel (that represents the node to be simulated) consists of three lag vectors **L**={(1,2),(2,1),(-1,-1)} corresponding to the relative locations of the three already simulated grid nodes.
   c) Define the data event $\mathbf{d}_n(\mathbf{x}, \mathbf{L}) = \{Z(\mathbf{x} + \mathbf{h}_1), \cdots, Z(\mathbf{x} + \mathbf{h}_n)\}$. It is a vector containing the values of the variable of interest at all the nodes of the neighborhood. In the example of Figure 1a, the data event is $\mathbf{d}_n(\mathbf{x}, \mathbf{L}) = \{0, 0, 1\}$.
   d) Define the search window in the TI. It is the ensemble of the locations **y** such that all the nodes $\mathbf{N}(\mathbf{y}, \mathbf{L})$ are located in the TI. The size of the search window is defined by the minimum and maximum values of the individual components of the lag vectors (Figure 1b).
   e) Randomly draw a location **y** in the search window and from that location scan systematically the search window. For each location **y**:
      i) Find the data event $\mathbf{d}_n(\mathbf{y}, \mathbf{L})$ in the training image. In Figure 1c, a random grid node has been selected in the search window of the TI. The data event is retrieved and is found to be $\mathbf{d}_n(\mathbf{y}, \mathbf{L})$={1,0,1}.
      ii) Compute the distance $d\{\mathbf{d}_n(\mathbf{x}, \mathbf{L}), \mathbf{d}_n(\mathbf{y}, \mathbf{L})\}$ between the data event found in the SG and the one found in the TI. The distance is computed differently for continuous or discrete variables. Therefore we will describe this step more in detail later in the paper.
      iii) Store **y,** $Z(\mathbf{y})$ and $d\{\mathbf{d}_n(\mathbf{x}, \mathbf{L}), \mathbf{d}_n(\mathbf{y}, \mathbf{L})\}$ if it is the lowest distance obtained so far for this data event.
      iv) If $d\{\mathbf{d}_n(\mathbf{x}, \mathbf{L}), \mathbf{d}_n(\mathbf{y}, \mathbf{L})\}$ is smaller than the acceptance threshold *t*, the value $Z(\mathbf{y})$ is sampled and assigned to $Z(\mathbf{x})$. This step is illustrated in

Figure 1d. In that case, the current data event in the TI matches exactly the data event in the SG. The distance is zero and the value $Z(\mathbf{y})=1$ is assigned to the SG (Figure 1e).

v) If the number of iterations of the loop i-iv exceeds a certain fraction of the size of the TI, the node $\mathbf{y}$ with the lowest distance is accepted and its value $Z(\mathbf{y})$ is assigned to $Z(\mathbf{x})$.

The definition of the data event consisting in considering the $n$ closest informed grid nodes is very convenient as it involves that the radius of the data events decreases as the density of informed grid nodes becomes higher. This natural variation of the data events size has the same effect as multiple-grids (Strebelle, 2002), and makes their use unnecessary. Figure 2 illustrates the decrease of the data events radius with neighborhoods defined by the four closest grid nodes.

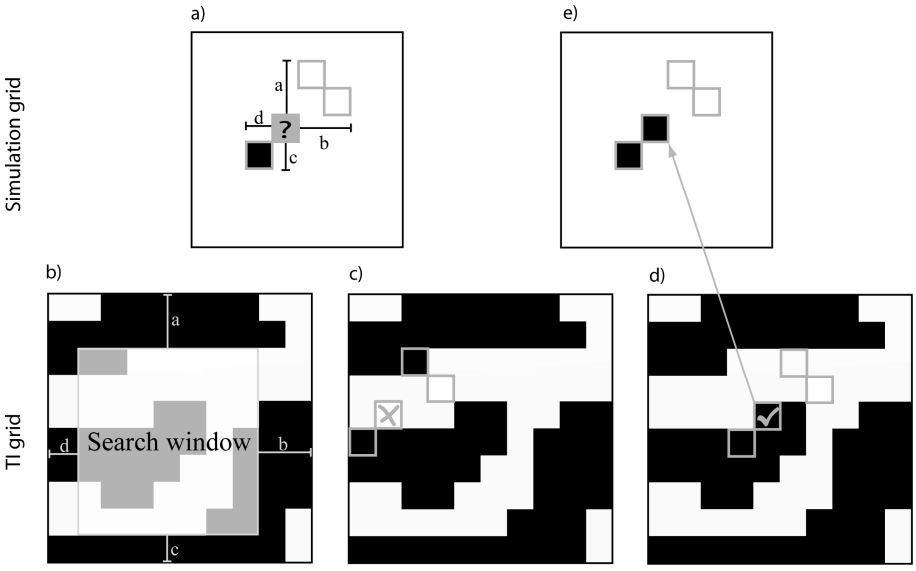



**Figure 1** Illustration of the Direct Sampling (DS) method. a) Define the data event in the simulation grid. The question mark represents the node to be simulated. The 2 white and the black pixels represent nodes that have been previously simulated. b) Define a search window in the TI grid by using the dimensions a,b,c,d of the data event. c) Linearly scan the search window starting from a random location until d) the simulation data event is satisfactorily matched. e) Assign the value of the central node of the first matching data event to the simulated node.

In the proposed method, the concept of a distance between data events $d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}$ is extremely powerful because it is flexible and can be adapted to the simulation of both continuous and categorical attributes. For categorical variables, we propose to use the fraction of non-matching nodes in the data event, given by the indicator variable *a* that equals 0 if two nodes have identical value and 1 otherwise:

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}=\frac{1}{n}\sum_{i=1}^{n}a_i \quad \in [0,1], \quad where\ a_i=\begin{cases}0 & if\ Z(\mathbf{x}_i)=Z(\mathbf{y}_i)\\ 1 & if\ Z(\mathbf{x}_i)\neq Z(\mathbf{y}_i)\end{cases}. \tag{2}$$

This measure of distance gives the same importance to all the nodes of the data event regardless of their location relative to the central node. It may be preferable to weight equation (2) according to the norm of the lag vector $\mathbf{h}_i$ using a power function of order $\delta$:

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}=\frac{\sum_{i=1}^{n}a_i\|\mathbf{h}_i\|^{-\delta}}{\sum_{i=1}^{n}\|\mathbf{h}_i\|^{-\delta}} \quad \in [0,1], \quad where\ a_i=\begin{cases}0 & if\ Z(\mathbf{x}_i)=Z(\mathbf{y}_i)\\ 1 & if\ Z(\mathbf{x}_i)\neq Z(\mathbf{y}_i)\end{cases}. \tag{3}$$

Specific weights can be defined if some of the data event nodes are conditioning data, as described in Zhang, et al. (2006). This can be used to enforce more patterns consistency in the neighborhood of conditioning data, or to give less importance to data presenting measurement errors. For all examples presented in this paper, we did not define specific weights for conditioning data. We also used $\delta = 0$ (i.e. all nodes of the data event have the same importance), which generally gives good results. Nevertheless, adjusting $\delta$ may allow obtaining images more representative of the TI while reducing CPU time.

Kriging weights could be used here instead of power distance weighting, but this return to multi-Gaussian techniques would involve tedious adjustment of covariance functions, which may have little significance if the training image is highly structured. Moreover, the CPU overburden involved in inverting a kriging matrix for each simulated node would be a high price to pay.

For continuous variables, we propose to use the distance:

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}=\sqrt{\sum_{i=1}^{n}\alpha_i\left[Z(\mathbf{x}_i)-Z(\mathbf{y}_i)\right]^2} \quad \in [0,1], \tag{4}$$

where

$$\alpha_i = \frac{\left\|\mathbf{h}_i\right\|^{-\delta}}{d_{\max}^2 \sum_{j=1}^{n} \left\|\mathbf{h}_j\right\|^{-\delta}}, \quad d_{\max} = \max_{y \in TI} Z(y) - \min_{y \in TI} Z(y). \tag{5}$$

The proposed distance is the square root of the weighted mean square differences between $\mathbf{d}_n(\mathbf{x})$ and $\mathbf{d}_n(\mathbf{y})$. In practice, the data event $\mathbf{d}_n(\mathbf{y})$ matching perfectly $\mathbf{d}_n(\mathbf{x})$ is often not found in the TI, especially for continuous variables. This is why an acceptance threshold *t* is introduced. When $d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}$ is smaller than *t,* the data event $\mathbf{d}_n(\mathbf{y})$ is accepted.

The numerator in $\alpha_i$ allows weighting the contribution of the data event nodes according to their distance to the central node. The denominator, although not needed for comparing distances between data events, is useful in practice to ensure that the distances are defined within the interval [0,1], making it easier to choose an appropriate acceptance threshold (for example, numerical tests have shown that 0.05 is a low threshold and 0.5 a high threshold, whereas it can be more tedious without normalization).

The distances proposed above do not pretend to be exhaustive and to be adapted to all possible situations. Other distances than the ones proposed above can be developed. For example, an alternative to (4) for continuous variables could be the normalized pair wise Manhattan distance,

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\} = \frac{1}{n}\sum_{i=1}^{n} \frac{\left|Z(\mathbf{x}_i) - Z(\mathbf{y}_i)\right|}{d_{\max}} \in [0,1]. \tag{6}$$

The choice of the distance measure used to compare data events of the simulation and of the TI should be adapted to the nature of the variable to simulate. For example, using distance (4) for the simulation of a categorical variable such as lithofacies would induce order relationships between the facies (i.e. facies 1 would be closer to facies 2 than to facies 3), which is conceptually wrong because facies codes are arbitrarily attributed.

In the examples we show how custom distances can be defined for specific problems (see below for a distance adapted to non-stationary cases). The concept of distance between data events is very general and therefore makes the method extremely flexible for a wide range of applications.

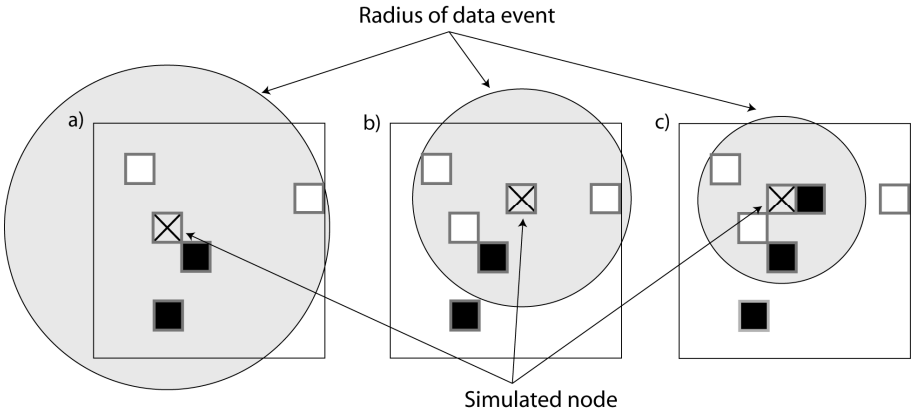



**Figure 2** Illustration of the natural reduction of the data events size. The neighborhoods for simulating three successive grid nodes a), b) and c) are defined as the 4 closest grid nodes. As the grid becomes more densely informed, the data events become smaller.

The quality of the pattern reproduction in the generated images depends on the size of the neighborhoods, the value of the acceptance threshold and the fraction of the TI that can be scanned for the simulation of each node. Certain settings of these parameters can be expensive in terms of CPU time. However, CPU burden can be alleviated using parallelization. Parallelizing the DS algorithm is straightforward on shared memory machines: each CPU performs the search in a limited portion of the TI. Our experience showed that this parallelization technique, using the OpenMP libraries, is very efficient in terms of speed-up. On a dual-core processor, the code runs about 1.9 times faster on two cores than on one, using various test cases. Moreover, recent parallelization strategies using Graphics Processing Units (GPU) let us hope for much shorter computation times. Nevertheless, even without parallelization, DS takes about the same time as traditional multiple-points simulators to obtain images of a similar quality.

## 4. Simulation of a continuous variable

Flow and transport simulators deal with continuous properties, such as hydraulic conductivity, storativity, porosity, etc. However, categorical images generation methods are usually preferentially used because they allow obtaining more realistic connectivity patterns. The facies obtained are then populated with continuous properties using other geostatistical techniques (Caers, 2005). DS allows bypassing

this 2-steps approach by directly generating continuous variables fields presenting realistic connectivity patterns.

Figure 3 shows a simulation using a TI borrowed from Zhang, et al. (2006), with continuous variable and high connectivity of the low values. The TI (Figure 3a) and the simulation (Figure 3b) have the same size of 200 by 200 grid nodes. Distance (4) was used in the DS simulation. Conditioning data are 100 values taken in the TI and located at random positions in the simulation. This ensures that the conditioning data are not spatially coherent with the model but belong to the univariate marginal distribution. Despite this situation, the DS algorithm produces realizations that are consistent with the TI (high connectivity of the low values) and satisfactorily respect the conditioning data. Figure 3c shows the histogram reproduction. Note that a unilateral path was used here (Pickard, 1980; Daly, 2004). Conditioning to data is possible with the unilateral path; this is accomplished by using large data events (80 nodes) including distant data points, which was not easily feasible with traditional multiple-points methods.

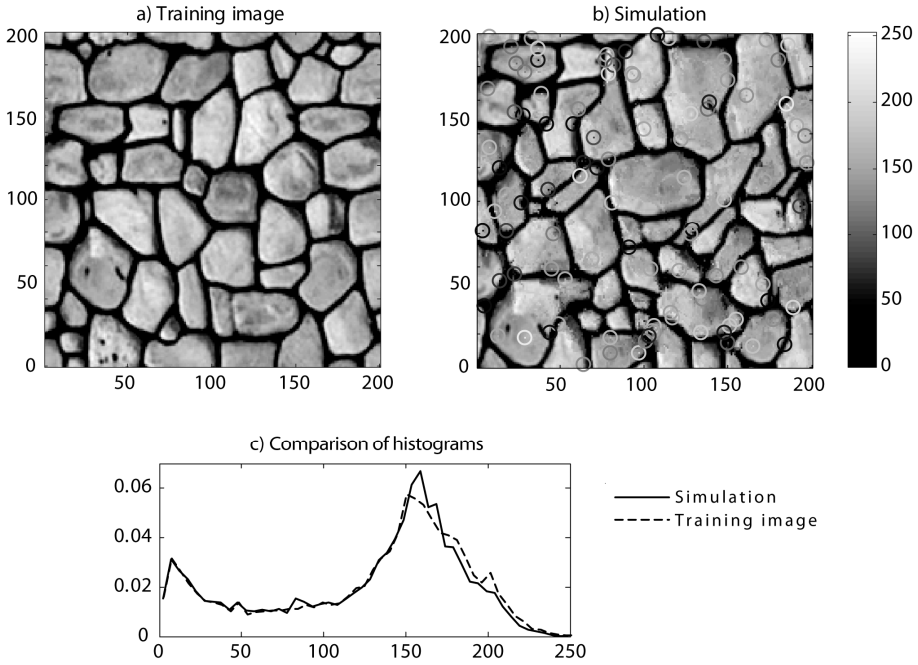



**Figure 3** Illustration of the method using a continuous variable. a) Training image with continuous variable. b) One simulation using the unilateral path with 100 randomly located conditioning data ($n$=80, $t$=0.01). Positions of

conditioning data are marked by circles whose colors indicate the values of the data. c) Comparison of the histograms.

The ability to reproduce connectivity patterns is of utmost importance for transport problems. Failure to reproduce the connectivity of high values of hydraulic conductivity can lead to serious underestimation of contamination rates, for example when modeling landfill leakages. To illustrate the effect of connectivity on transport, the result of solute migration was compared on two synthetic aquifers, the first one being generated using Direct Sampling with a continuous variable, and the second one generated with Sequential Gaussian Simulation (Journel and Isaaks, 1984).

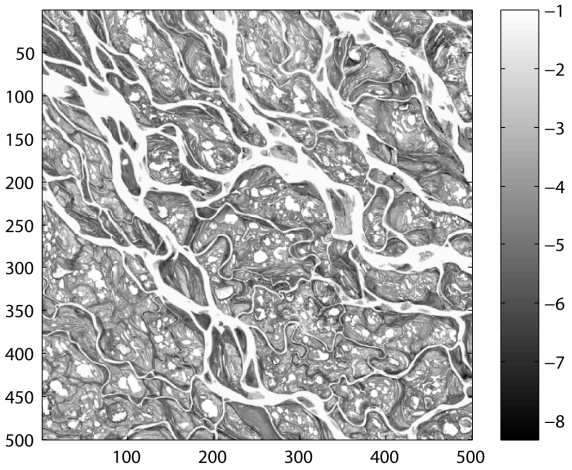



**Figure 4** $Log_{10}$ hydraulic conductivity values based on the color values of a satellite image of the Lena Delta (Landsat 7 image, USGS/EROS and NASA Landsat Project). This image constitutes the spatial model for generating the synthetic aquifers in the solute migration problems (either as Training Image for DS or for inferring the variogram and histogram for SGS).

The desired spatial model for the synthetic contaminated aquifer is shown in Figure 4. The image is a satellite photograph of the Lena Delta (Russian Federation) where complex structures are present at different scales. Large channels are continuous from one side of the image to the other. Smaller, more tortuous channels define connectivity of high values at a local scale. Small lakes are also present and constitute bodies of a completely different nature that coexist at the same scale than the small channels. The color values of this image have been converted to hydraulic conductivities spanning between $10^{-1}$ and $10^{-8}$ m/s. The two heterogeneous hydraulic

conductivity fields (size 250 by 250 m) have been generated with this spatial model in mind. Figure 5a is an unconditional DS simulation with Figure 4 as TI, with distance (4), $n$=30 and $t$=0.01. Figure 5b is an unconditional SGS simulation using a variogram model adjusted on the experimental variogram of Figure 4. As the variable is not multi-Gaussian, a histogram transform was performed with full knowledge of the reference histogram. Therefore, Figure 5b has the same histogram and variogram as Figure 4.

Both synthetic fields share similar spatial properties when one only looks at the histograms (Figure 5c) and the variograms (Figure 5d). However, when visually comparing the resulting fields, the DS simulation obviously shows more realistic connectivity patterns (i.e. closer than what is observed in Figure 4).

A simple flow and transport problem was set for each hydraulic conductivity field, with permanent flow and transient transport regime. Constant head boundaries were set on both sides of the domain, with $H$=1 on the western side, and $H$=0 on the eastern side, and no flow boundaries on the North and South limits. A constant concentration was defined on the western side of the domain, with $c$=1 mg/l, which can represent a leaking landfill.

Figure 5e shows the evolution of the average concentration observed on the outflow boundary of the model (eastern side). Because the connectivity patterns are not the same in both models, the contaminant breakthrough curves are very different. Whereas high concentrations have crossed the entire domain after about 250 hours in the first case (Direct Sampling), they do not reach the same level after 5000 hours in the second case (SGS). The differences in transport behavior are further illustrated by Figure 5f and Figure 5g that display the distribution of contaminant after 250 hours. The influence of the large channels and their connectivity is very visible in the first case. In the second case, even though fingering occurs due to the presence of heterogeneity, the high conductivity values are not connected enough to drive a fast solute transport.

This example shows that honoring the histogram and the variogram is not enough to obtain realistic connectivity patterns, and emphasizes the need for methods such as DS, able to deal with low-entropy and non-multi-Gaussian structures.

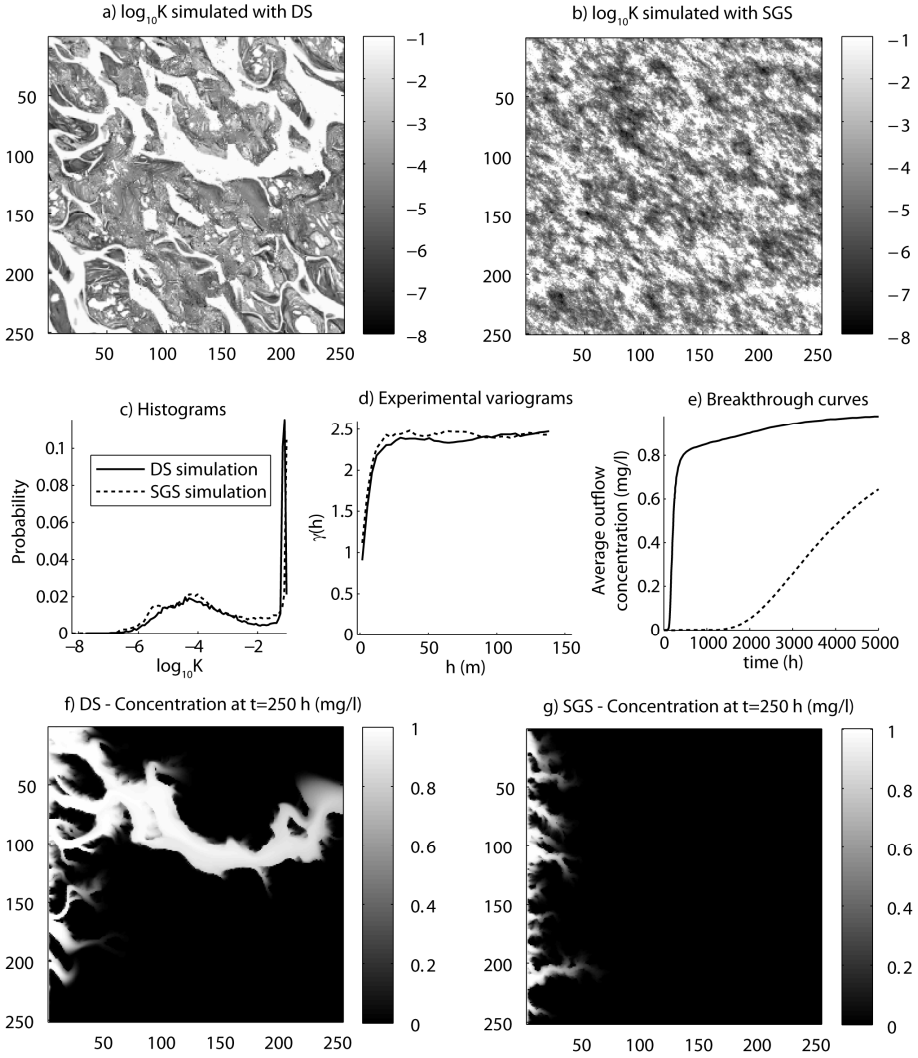



**Figure 5** Influence on the connectivity patterns on transport behavior. a) One simulation obtained with Direct sampling. b) One simulation generated using SGS. c) Comparison of histograms of simulated values. d) Comparison of experimental omnidirectional variograms. e) Comparison of contaminant breakthrough curves on the outflow boundary. f) and g) comparison of the contaminant distribution at t=250 hours.

## 5. Multivariate case

Contrarily to existing multiple-points simulation techniques, DS is not limited by the dimension of the data events because there is no need to store their occurrences. This allows defining the data events through a set of variables that can be simulated jointly or used for conditioning following the same principle as co-simulation (it may be collocated or not). The training image is a multivariate field comprising $m$ variables $Z_1(\mathbf{x}),\cdots, Z_m(\mathbf{x})$. Such multivariate fields are presented as "vector images" by Hu and Chugunova (2008). Accounting for multiple-points correlations between variables means to respect cross-correlations between all combinations of nodes within multivariate data events. The conditional cumulative density function (1) for the variable $Z_k$ is then expressed as

$$F_k(z,\mathbf{x},\mathbf{d}^1_{n_1},\cdots,\mathbf{d}^m_{n_m}) = \text{Prob}\left\{Z_k(\mathbf{x}) \le z \mid \mathbf{d}^1_{n_1},\cdots,\mathbf{d}^m_{n_m}\right\}, \quad k=1,\cdots,m. \tag{7}$$

Each variable $Z_k$ involved in the multivariate analysis can have a different neighborhood and a specific data event $\mathbf{d}^k_{n_k}(\mathbf{x},\mathbf{L}^k) = \left\{Z_k(\mathbf{x}+\mathbf{h}^k_1),\cdots,Z_k(\mathbf{x}+\mathbf{h}^k_{n_k})\right\}$. This involves that the number $n_k$ of nodes in the data event of each variable can be different, as well as the lag vectors $\mathbf{L}^k$. To simplify the notation, we just extend the previous concept of data event to the multivariate case: here the data event $\mathbf{d}_n(\mathbf{x})$ is the joint data event including all the individual data events $\mathbf{d}_n(\mathbf{x}) = \left\{\mathbf{d}^1_{n_1}(\mathbf{x},\mathbf{L}^1),\cdots,\mathbf{d}^m_{n_m}(\mathbf{x},\mathbf{L}^m)\right\}$. The distance between a joint data event found in the simulation and one found in the TI is defined as a weighted average of the individual distances defined previously:

$$d\left\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\right\} = \sum_{k=1}^{m} w_k\, d\left\{\mathbf{d}^k_{n_k}(\mathbf{x},\mathbf{L}^k),\mathbf{d}^k_{n_k}(\mathbf{y},\mathbf{L}^k)\right\} \in [0,1], \text{ with } \sum_{k=1}^{m} w_k = 1, \text{ and } w_k \ge 0 \cdot \tag{8}$$

The weights $w_k$ are defined by the user. This allows accounting for the fact that the pertinent measure of distance may be different for each variable. Multivariate simulations are performed using a single (random) path that visits all components of vector $Z$ at all nodes of the SG.

Figure 6 shows an example of a joint simulation of two variables that are spatially correlated by some unknown function. For this synthetic example, the TI for variable 1 (Figure 6a) is a binary image representing a channel system (Strebelle, 2002). The TI for variable 2 (Figure 6b) was obtained by smoothing variable 1 using

a moving average with a window made of the 500 closest nodes, and then adding an uncorrelated white noise uniformly distributed between 0 and 0.5. This secondary variable could represent the resistivity map corresponding to the lithofacies given by variable 1. The result is a bivariate training image where variables 1 and 2 are correlated via a multiple-points relationship. Figure 6c and Figure 6d show one unconditional bivariate simulation using the TI described above. The categorical variable 1 uses distance (3) and the continuous variable 2 uses distance (4). The multiple-point correlation relating both variables is well reproduced, both visually and in terms of cross-variograms (Figure 6e), which is a measure of two-points correlation. Note that addressing correlations between categorical and continuous variables is usually awkward. The scatter plot depends on the facies numbering (which is arbitrary) and correlation factors are meaningless. Here, DS is able to reproduce multiple-points correlations in a straightforward manner, including statistical parameters that are much more complex than the scatter-plot (e.g. cross-variograms).

Problems traditionally addressed by including exhaustively known secondary variables (e.g. Mariethoz, et al., In Press) are particular cases of the multivariate DS approach. Whereas existing MP methods consider only the secondary variable at the central node $\mathbf{x}$, DS accounts for complex spatial patterns of the secondary variable because multiple-points statistics are considered for both primary and secondary variables.

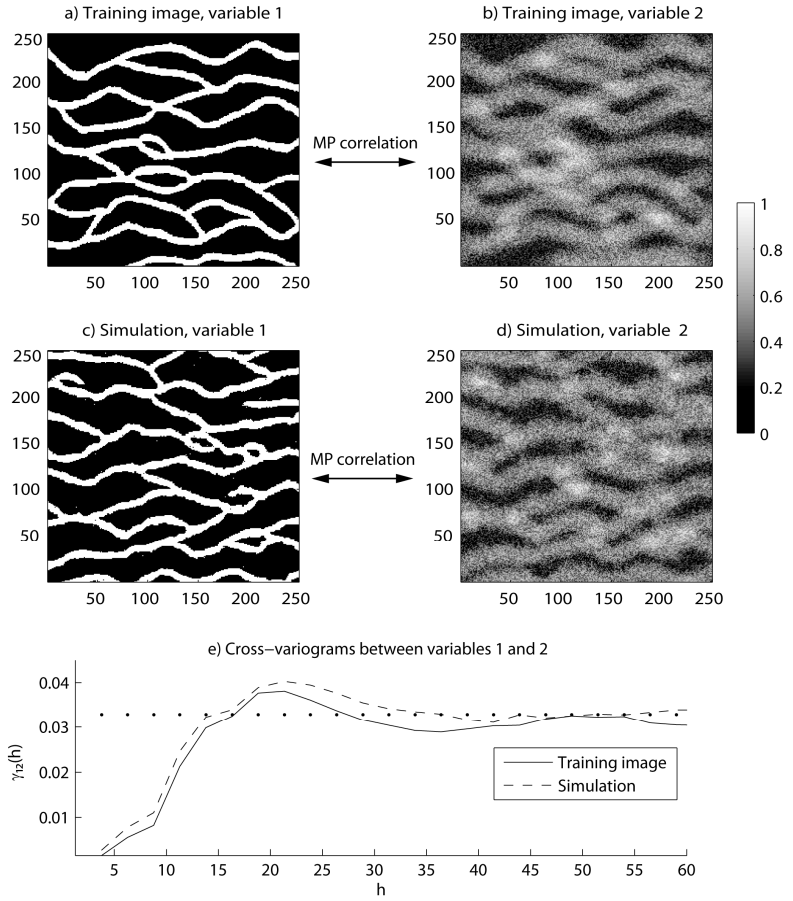



**Figure 6** Joint simulation of two variables ($n_1$=30, $n_2$=30, $t$=0.01, $w_1$=0.5, $w_2$=0.5). a) and b) The bivariate training image, with a complex multiple-points correlation. c) and d) One resulting bivariate simulation, where the MP correlation is reproduced. e) Cross-variograms reproduction. Note that no variogram adjustment was necessary.

When one (or several) of the joint variables is already known, DS uses such information as indirect conditioning data (secondary variable) guiding the simulation of the other variables (primary variables) and then reducing uncertainty. In this situation, the exercise is to reconstruct the primary reference field knowing only the secondary variable and the experimental relation between primary and secondary variables, which is given by the bivariate TI. For illustration, consider Figure 6a and Figure 6b as bivariate TI and Figure 7b as auxiliary variable for the

simulation grid. Figure 7b was obtained as follows. First, Figure 7a was generated with an univariate unconditional simulation using Figure 6a as TI. Then, Figure 7b was computed from Figure 7a, applying a moving average followed by addition of a white noise. Hence, the aim is to reconstruct the reference field Figure 7a from Figure 7b and the multiple-points correlation given by the bivariate TI (Figure 6a and Figure 6b), using multivariate DS.

Figure 7c displays one realization of the primary variable, conditional to the exhaustively known secondary variable (Figure 7b). No conditioning data are available for the primary variable. The features of the reference field are correctly inferred from the information contained in the secondary variable, as shown in Figure 7d, where the reference (Figure 7a) and the simulation (Figure 7c) are superposed. In Figure 7e, the mean of 100 simulations is presented. In average, the channels are correctly located when compared to the reference.

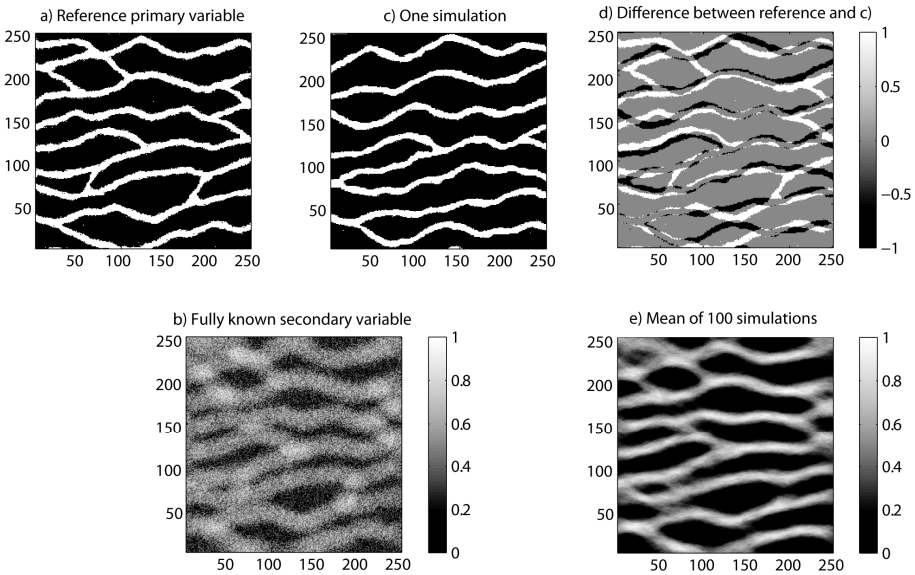



**Figure 7** The use of a secondary variable to guide the simulation of a primary variable. a) The reference primary variable, obtained with an univariate unconditional simulation using Figure 6a as TI. b) The reference secondary variable computed by transformations of the primary variable (see text for details). The bivariate training image a) and b) describes the MP relationship between primary and secondary variables. c) One multivariate simulation generated using the fully known secondary variable b) as conditioning data

($n_1$=30, $n_2$=30, $t$=0.01, $w_1$=0.5, $w_2$=0.5). d) Superposition of one simulation and the reference. e) Mean of 100 simulations.

A real application example would be a case where a geophysical survey provides an exhaustive map of resistivity (secondary variable), but where the primary variable to be characterized is hydraulic conductivity (for example, Langsholt, et al., 1998 are confronted to a similar problem with GPR data). The relation between both variables can be complex and not necessary linear. This problem would require building a bivariate TI.

One possible methodology would consist in obtaining the secondary variable of the TI using a forward simulator, with as input the primary variable of the TI. In the synthetic example above, the secondary variable of the TI was obtained by moving average followed by addition of a white noise. A forward problem (e.g. simulating a field of resistivity values based on the known facies distribution) could be used instead of this rather simple transformation.

## 6. Dealing with non-stationarity

Geological processes are intrinsically non-stationary. The ability to address non stationarity is vital for the applicability of a geostatistical method in Earth Sciences. For existing MP methods, several techniques can be found in the literature to account for non-stationarity either of the TI or of the simulated field (Journel, 2002; Strebelle, 2002; Chugunova and Hu, 2008; De Vries, et al., 2009). Most of these techniques can be used with DS, but new possibilities are also offered by exploiting the specificities of DS.

One way of dealing with non-stationary TIs is to divide a non-stationary TI in stationary zones, each considered as a separate stationary TI (De Vries, et al., 2009; Boucher, In Press). The simulation domain is divided into zones, each corresponding to a specific TI. In the framework of traditional multiple-points statistics, using multiple TIs involves creating one data events catalogue per training image (Wu, et al., 2008). DS can accommodate this situation by scanning a different TI for each simulated zone. There are no limitations to the number of TIs and zones related to memory requirements because data events catalogues are not stored.

Non-stationarity related to local variations of facies proportions can be addressed by using soft probability maps to alter the local ccdfs. This soft probability indicates areas where certain facies are more likely to occur. The most common techniques for updating the local ccdf (1) are Bordley's formula (Bordley, 1982) and the tau

model (Journel, 2002). Such techniques cannot be directly applied to DS, because the local ccdf (1) is never computed, and therefore cannot be altered. Instead, it may be possible to obtain similar results by including a penalty term in the distance measure that would be proportional to the local probability of occurrence of a certain facies.

Addressing non-stationarity can be accomplished in many other ways. We illustrate below three possible methods to deal with non-stationarity when using Direct Sampling.

## 6.1. Addressing non-stationarity with specific distances

We discussed above how the distance measure should be chosen according to the nature of the variables at stake. Following this idea, we propose to forge distances adapted to non-stationary cases. An example of such custom distance measure is the pair-wise Euclidean distance relative to the mean of the data event:

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}=\left(\sum_{i=1}^{n}\alpha_i\left[\left(Z(\mathbf{x}_i)-\overline{Z}(\mathbf{x})\right)-\left(Z(\mathbf{y}_i)-\overline{Z}(\mathbf{y})\right)\right]^2\right)^{1/2}\in[0,1], \qquad (9)$$

with $\overline{Z}(\mathbf{x})=\frac{1}{n}\sum_{i=1}^{n}Z(\mathbf{x}_i)$. When a matching data event in found in the TI, the local mean of the SG data event is added to the value found in the TI. Therefore, the value $z(\mathbf{y})+\overline{Z}(\mathbf{x})$ is attributed to each simulated node. This distance involves that the data events are compared regarding their internal variations only, and not their actual values. This variation-based distance can be very useful when considering non-stationary phenomena. We illustrate this situation with the example depicted in Figure 8. The available training image (Figure 8a) is a multi-Gaussian field with zero mean and unit variance, resulting in minimum and maximum values of respectively -3.52 and 3.99. It was generated using an exponential variogram model, with ranges of 35 units along the x axis and 25 units along the y axis. Its size is 250 by 250 grid nodes. 100 conditioning data are available (Figure 8b), but they are not compatible with the training image, as their values span between a minimum of 99.55 and a maximum of 110.92, with a mean of 105.12. Moreover, these data show non-stationarity. Because the distance (9) is based on the variations of the values in the data event, it allows finding matches between the data events found in the data and the ones of the TI despite the difference in the range and the non-stationarity. The resulting simulations (one is shown in Figure 8c) display the same variable

range (minimum: 98.13, maximum: 111.72, mean: 104.87) and the same non-stationary behavior as the data, but also a spatial structure similar to what is found in the TI. In this case, non-stationarity can be seen as a locally varying mean, and therefore distance (9) can accommodate it well. If the non-stationarity was more complex, such as for example structures ranging from channels to lenses, this distance measure would not be appropriate.

This example shows that variation-based distance can be very useful in real cases where a conceptual model allows the geologist to provide a training image, but where the data indicate the presence of non-stationarity and inadequacy with the ranges given in the TI. Moreover, it emphasizes the flexibility offered by using distances between data events, which is one of the major advantages of the DS approach.

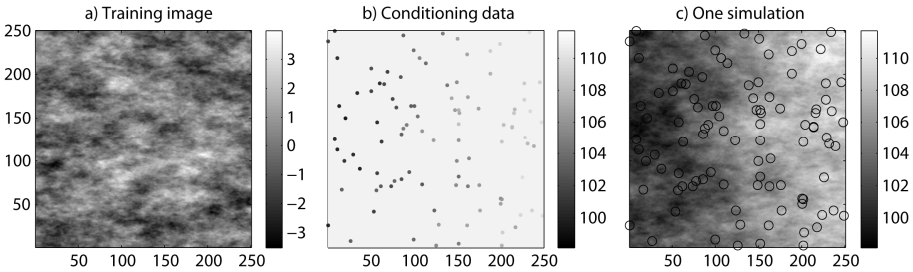



**Figure 8** Simulation using a variation-based distance. a) multi-Gaussian stationary training image. b) Non-stationary data set (100 points data), with values in a different range than those of the training image. c) One simulation with variation-based distance ($n$=15, $t$=0.01). Circles represent the location of the 100 conditioning data.

## 6.2. Addressing non-stationarity with transformation of data events

Traditional multiple-points simulation implementations such as *snesim* include the possibility to impose transformations on the structures found in the TI. This is done by first constructing the data events catalogue using a transformed template, and then simulating values with a non-transformed template (Strebelle, 2002). The most commonly implemented transformations are rotation and affinity (dilatation). This feature is very useful when the modeler has a single stationary training image and wants to use it for the simulation of non-stationary fields. If many different transformations have to be applied on the simulation grid, most approaches store as many data events catalogues. The DS approach also allows these transformations.

Simply scanning the TI with a transformed data event gives the same results as the traditional technique. Moreover, transformations are not defined by zones, but as a continuum, because the transformation can be different for each simulated node. In some cases, rotation or affinity may result in large data events that do not fit in the TI. In such cases, the data event nodes located outside of the TI are ignored until it becomes possible to scan the TI with this new, reduced data event.

Figure 9 shows an example of such transformation, with angle and affinity maps (Figure 9a and Figure 9b) defined by continuous variables. All angles between -180 and 180 degrees are represented, and the affinity ratios range from 1 at the center of the image to 0.4 in the corners (meaning that all structures are reduced to 40% of the size they have in the TI). The training image (Figure 9c) (Strebelle, 2002) is much smaller (250 by 250 nodes) than the simulation domain (1000 by 1000 nodes) and represents horizontal channels. This combined transformation (rotations + affinities) results in channels oriented in all directions and becoming thinner as they are located further away from the centre (Figure 9d).

## 6.3. Addressing non-stationarity with a secondary variable

Another situation addressed by using a secondary variable occurs when there is a need to use a non-stationary univariate training image (for example made from direct field observation or from a simulation of sedimentary processes). In this case, secondary variables can be used to model the non-stationarity as it was proposed by Chugunova and Hu (2008). Using multivariate DS, it can be accomplished in a straightforward manner by again using a bivariate TI with variable 1 being the variable of interest and variable 2 describing the non-stationarity of variable 1. Figure 10 illustrates this situation. The TI for variable 1 is a set of rotating channels such that their mean orientation makes an angle with the horizontal axis that changes as a function of the $X$ coordinate (Figure 10a). Describing this non-stationarity is thus simply mapping the $X$ coordinate of the TI grid (Figure 10b). Variable 2 has the same function on the SG grid as on the TI grid: it describes the non-stationarity of variable 1. This secondary variable is entirely informed on the SG and it consists of the $Y$ coordinate of each grid node. Having variable 2 exhaustively known on the simulation grid means that we know how the simulated channels should be oriented.

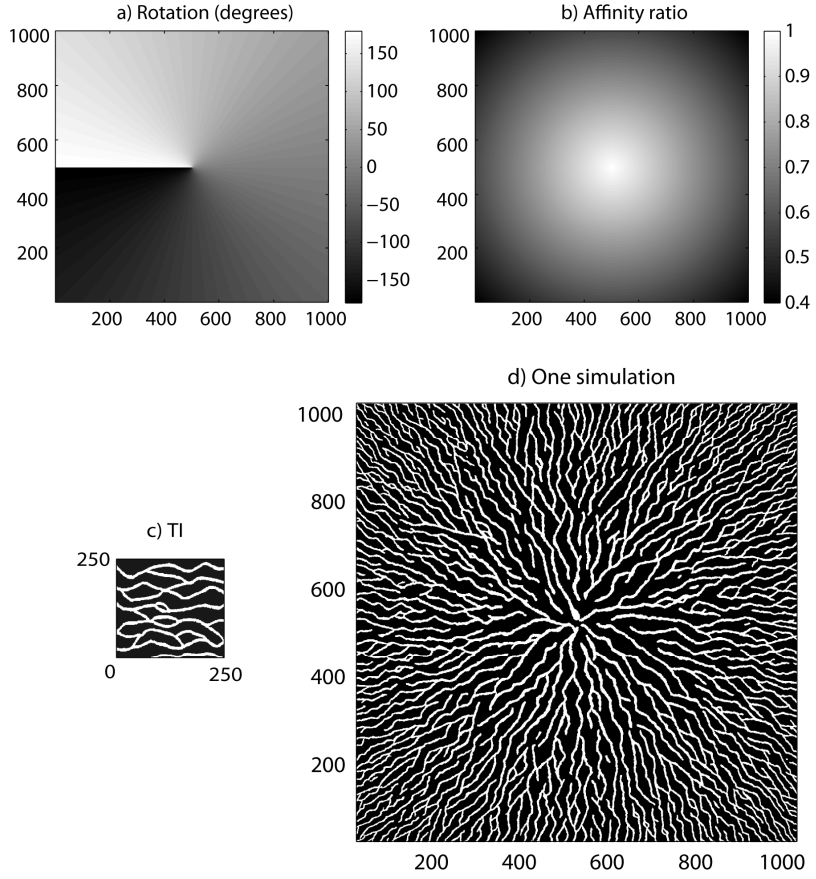



**Figure 9** Transformations of the data events. a) Rotation map. b) Affinity map. c) Stationary training image. d) Simulation with transformed data events ($n$=30, $t$=0).

Simulations (Figure 10c) have been generated using the bivariate training image and the exhaustively known variable 2 as conditioning data (Figure 10d). The neighborhoods are made of $n_1$=30 nodes for variable 1 and $n_2$=1 for variable 2, because a single node is enough to characterize the information carried by variable 2. If a more complex non-stationarity was to be modeled, larger neighborhoods should be used for variable 2. The weight of both variables are equal, with $w_1$=$w_2$=0.5. Similarly to the TI, bivariate simulations have horizontal channels at locations where variable 2 takes a value 0, vertical channels where it takes a value 1, and channels oriented SW-NE where it takes a value 0.5. However, the simulated

orientations are not the same as the ones of the TI because both maps of variable 2 are different.

Although this is a simple demonstration, the use of a continuous secondary variable to describe non-stationarity allows accounting for very rich types of non-stationarity such as a change in the type of structures encountered. Using large neighborhoods for the secondary variable allows accounting for complex non-stationarity.

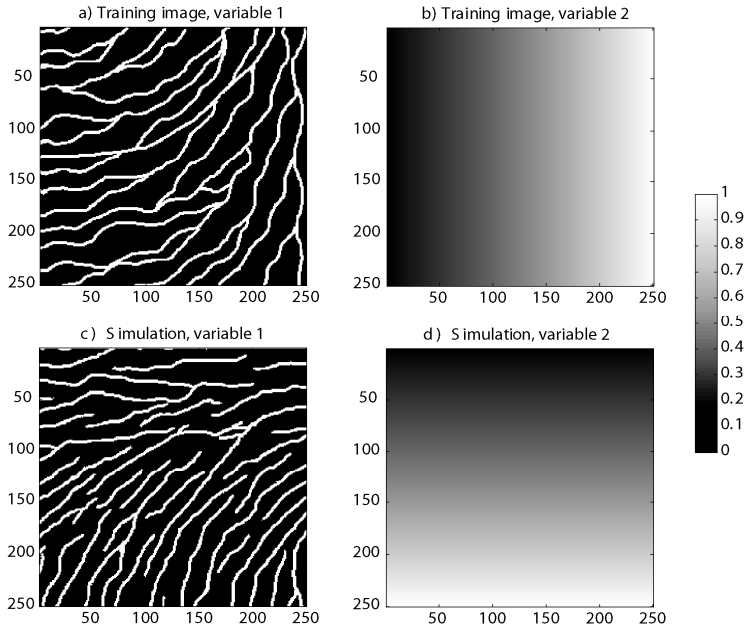



**Figure 10** The use of a secondary variable to model non-stationarity a) Variable 1 of non-stationary training image. b) Correlated joint variable describing the non-stationarity of variable 1 in training image. c) Resulting simulation for variable 1 ($n_1$=30, $n_2$=1, $t$=0.01, $w_1$=0.5, $w_2$=0.5). d) Correlated joint variable (exhaustively known) describing the non-stationarity of variable 1 in simulation.

## 7. Improving pattern reproduction

Accurate pattern reproduction can be jeopardized when the data event cannot be found in the TI. This problem is common to all multiple-points simulations methods, and is more acute when a random path is used in the simulation grid. In traditional

multiple-points simulation algorithms, this issue is usually dealt with by dropping the neighbor node that is the farthest away from the central node, and by performing a search in the data events catalogue for this new, reduced pattern (Strebelle, 2002). The main drawback of this procedure is that it induces a degradation of the pattern reproduction by artificially reducing the template size for the computation of the ccdf (1). Such degradation can lead to a lack of spatial continuity of the simulated structures (such as channels). Several authors have proposed methods to improve patterns reproduction. Strebelle and Remy (2005) locate the nodes that were simulated using a reduced neighborhood, and re-simulate the dropped neighbors at the end of each multi-grid step. This method does not remove all the inconsistencies in the simulated patterns, but performs additional simulation attempts with updated neighborhoods. As problematic values are temporarily accepted (until the entire multi-grid is simulated), they propagate inconsistencies to nodes that are simulated later. Therefore, if a node is successfully re-simulated, it is not a guarantee that all its neighbors are consistent between each other. Another algorithm, proposed by Stien et al. (2007), does not temporarily accept values generating conflicts, but deletes the problematic nodes in the neighborhood. At the end of a multi-grid level, these nodes are simulated. The process is iterative and needs specific parameters to ensure convergence. Although this method avoids the propagation of inconsistencies by deleting them, it does not resolve the problem of the origin of these problematic patterns. Indeed, inconsistencies exist because other nearby problematic patterns occurred previously in the simulation process. In our opinion, the only way to deal with this problem is to immediately address the entire cascade of causes at the origin of problematic patterns.

In the context of simulations using a unilateral path (Daly, 2004), Suzuki and Strebelle (2007) developed the Real-Time Post-Processing Method (RTPP), that walks back the unilateral path when problematic neighborhoods are encountered, and re-simulates the most recent nodes until the produced patterns satisfactorily match the ones of the TI. The limits of this method are that it is applicable to the first stage of the simulation only (the first multi-grid), and only when using a unilateral path. Therefore, like all simulation methods using the unilateral model, it suffers from difficulties to honor conditioning data. Nevertheless, this method has the advantage of correcting all inconsistencies because it re-simulates the neighborhoods of the problematic nodes, and not only the problematic nodes themselves. As inconsistencies are re-simulated immediately, it avoids propagation to their neighbors.

In this context, we propose a new algorithm, the syn-processing, aimed at improving patterns reproduction. It is generic enough to be applicable to any type of path and with or without multi-grids. As the RTPP, it re-simulates values as soon as inconsistencies are met. It is based on the idea that when an inconsistent pattern (with respect to the TI) is found, it is because other inconsistencies occurred upstream in the simulation process. Therefore, before re-simulating problematic nodes, their neighborhoods also need to be (at least partially) re-simulated. If inconsistencies appear during this re-simulation, further re-simulation needs to be undergone. Hence, the algorithm is of a recursive nature.

The syn-processing algorithm consists in the following steps at each simulated node $\mathbf{x}$:

1. Check if the simulated value $Z(\mathbf{x})$ is acceptable. The acceptance criterion can be a minimum number of dropped neighbor nodes in the framework of classical multiple-points implementation. In the case of DS, the criterion is that the minimum distance $d\{\mathbf{d}_n(\mathbf{x}), \mathbf{d}_n(\mathbf{y})\}$ found is below a threshold.
2. If the criterion is not met, the simulation of $Z(\mathbf{x})$ is postponed and one of its neighbors $\mathbf{N}^{-1}(\mathbf{x})$, taken among those that do not belong to the original set of conditioning data, is re-simulated taking into account the same criterion. For the simulation of the node $N^{-D}(\mathbf{x})$:
    a. If the criterion is not met for $Z\{N^{-D}(\mathbf{x})\}$, delete it and re-simulate one of its neighbors, $N^{-(D+1)}(\mathbf{x})$.
    b. If the criterion is met for $Z\{N^{-D}(\mathbf{x})\}$, accept this value and try simulating $Z\{N^{1-D}(\mathbf{x})\}$.

    Note that $D$ is the number of deleted nodes for the initial node $\mathbf{x}$. To ensure convergence, a maximum allowed number of deletions must be set.

Syn-processing can sometimes delete and re-simulate the same nodes in a cyclic manner. Such cycles are a waste of time as they do not improve the simulation. This can be avoided by keeping a record of all deletions. Before each deletion, the analysis of this record allows finding if the present state of the SG already occurred in the past. If it is the case, another random neighbor is chosen in order to break the cycle.

Tests showed that syn-processing efficiently reproduces patterns, as well as conditioning to local data. As the algorithm is recursive, CPU time can be adversely

affected depending on the criterion to accept a simulated value. If the criterion is very strict (such as $t$=0 for a continuous variable) and if the maximum allowed number of iterations is very large, convergence can be compromised. On the other hand, improving pattern reproduction increases the global coherence of the simulation with respect to the TI. It becomes then easier to find matching data events in the TI, thus making the scan process faster for the remaining nodes. In certain cases, syn-processing can even reduce simulation time up to a factor 2 while improving simulations quality. Moreover, tests showed that performing syn-processing only at the beginning of the simulation is sufficient to obtain better reproduction of large scale features and general connectedness of the simulated structures, as compared to simulations without syn-processing. Therefore, a compromise has to be found between the different parameters governing the simulation, in order to obtain optimal results at the lesser CPU cost.

Note that syn-processing was used when generating the simulation examples presented in Figure 7a and Figure 10. For comparison, note the difference of continuity between Figure 6c, where no syn-processing was used, and Figure 7a that was generated using syn-processing. The latter figures display channels that reproduce better the sinuosity and the connectivity of the channels found in the TI.

## 8. Discussion and conclusion

In this paper, we presented the Direct Sampling (DS), a simulation method that has several practical and computational advantages compared to traditional multiple-points techniques. Additionally, a recursive syn-processing algorithm was developed that allows enhancing simulations by forcing reproduction of the patterns found in the TI.

By using distances between data events, DS has the flexibility to accommodate training images that can be either categorical or continuous, uni- or multivariate, stationary or non-stationary. This can be invaluable when realistic geological structures must be characterized. It can be the case for hydrogeological problems influenced by heterogeneity and connectivity of the geological structures, such as contaminant migration modeling.

The multivariate framework offered by DS opens new perspectives for the integration of different data types in geological models. By accounting for multiple-points correlations, it can exploit experimental relationships between variables of different nature, such as between categorical and continuous variables. A possible hydrogeological application could be to link the occurrence of lithofacies to indirect

measurements obtained by geophysical surveys. We believe other uses of this feature can be envisioned. In hydrology, categorical variables are widely employed to characterize interpreted information (e.g. soil type, land cover category, vulnerability class), whereas continuous variables are issued directly from measurements (e.g. porosity, concentration, recharge rate). DS addresses the challenging issue of integrating both types of variables.

DS may also be able to deal with classes of problems and fields of application that were usually not addressed by multiple-points methods. Applications can be envisioned in domains using digital images or even for the cosimulation of time series such as synthetic rainfall or financial indices.

In addition to the wide spectrum of potential applications, DS also has computational advantages. Compared to traditional Multiple-Points methods, it massively reduces memory usage because no catalogue of data events needs to be stored. This implies that the size of the neighborhood is not limited by memory considerations. The data event can be spread on many different variables, allowing to perform multivariate simulations of variables presenting complex multiple-points correlations.

Because multiple-points statistics are not stored, DS does not need a fixed geometry of the data events. The shape of the data event can change at each simulated node and so does the search window. Hence, the data events are always adapted to the simulation path. The size of the data events is only limited by the size of the TI, and is controlled by a maximum number $n$ of nodes. In certain cases, it can be useful to limit the radius of the data events, for example, when considering non-stationary variables, to avoid capturing non-stationarity within the data events. It is also useful if the simulation is larger than the training image. In this case, very large data events can result in very small search windows, leading to a bias towards reproducing the statistical properties of a small central portion of the TI.

A related advantage of the DS approach is that multi-grids (a step-like decrease in the template dimension) are replaced by a progressive (linear) decrease of the size of the data event as a function of the density of simulated nodes. It ensures that structures of all sizes are present in the simulation. Abandoning multi-grids avoids problems related to the migration of conditioning data on coarse multi-grid levels. By avoiding multi-grids, DS is easy to implement, easy to parameterize and has no problems accommodating large datasets.

A very important point is that DS is not prohibitive in CPU time, with performances comparable to existing methods. This good performance is possible because the algorithm searches only a single matching data event and, therefore, the

whole TI often does not need to be scanned. There is a tradeoff between CPU time and the quality of the generated images, controlled by parameters such as the size of the neighborhoods, the value of the acceptance threshold and the fraction of the TI that can be scanned for the simulation of each node. However, using parallelization allows easily increasing the performance of DS.

The concept of acceptance threshold in the distance between data events put in question the model validity. Setting this threshold above a value 0 means to authorize inadequacies between the model (the TI) and the simulation. Our opinion is that at one point or another of the simulation process, one has to admit some discrepancy between the simulation and the model. Otherwise, the simulation is very likely to be a simple copy (or a patch) of the TI.

A possible direction for further research is the completion of partially informed images. Instead of scanning a training image for a specific data event, it is possible to scan a dataset, provided that it is rich enough to contain a variety of patterns. This could allow abandoning the use of a training image when large datasets are available. As indicated by Ortiz and Deutsch (2004), the model information contained by the data is based on the true field, which is probably better than a training image derived from interpretation. This is a reminder of the idea behind variogram fitting: to adjust the model on the field data.

## 9. References


Aitokhuehi, I., and Durlofsky, L. J. (2005), *Optimizing the performance of smart wells in complex reservoirs using continuously updated geological models*, Journal of Petroleum Science and Engineering, 48, 3-4, 254-264.

Alcolea, A., Renard, P., Mariethoz, G., and Bretone, F. (2009), *Reducing the impact of a desalination plant using stochastic modeling and optimization techniques*, Journal of Hydrology, 365, 3-4, 275-288.

Arpat, B., and Caers, J. (2007), *Conditional Simulations with Patterns*, Mathematical Geology, 39, 2, 177-203.

Bordley, R. E. (1982), *A multiplicative formula for aggregating probability assessments*, Management Science, 28, 10, 1137-1148.

Boucher, A. (In Press), *Considering complex training images with search tree partitioning*, Computers & Geosciences, 35, 6, 1151-1158.

Caers, J., Strebelle, S., and Payrazyan, K. (2003), *Stochastic integration of seismic data and geologic scenarios: a West Africa submarine channel saga*, The Leading Edge, 22, 3, 192-196.

Caers, J. (2005), *Petroleum Geostatistics*, Society of Petroleum Engineers.

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., and Luit, J. (2005), *Inverse problem in hydrogeology*, Hydrogeology Journal, 13, 206-222.

Christakos, G. (2004), *A sociological approach to the state of stochastic hydrogeology*, Stoch Envir Res and Risk Ass, 18, 274–277.

Chugunova, T., and Hu, L. (2008), *Multiple-Point Simulations Constrained by Continuous Auxiliary Data*, Mathematical Geosciences, 40, 2, 133-146.

Dagan, G. (1976), *Stochastic conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media - Comment*, Water Resour. Res., 12, 567-567.

Dagan, G. (1986), *Statistical theory of groundwater flow and transport: pore to laboratory, 25 laboratory to formation, and formation to regional scale*, Water Resour. Res., 22, 120S-134S.

Daly, C. (2004), *Higher order models using entropy, Markov random fields and sequential simulation*, paper presented at Geostatistics Banff 2004, Kluwer Academic Publisher, Banff, Alberta.

De Marsily, G., Delay, F., Gonçalvès, J., Renard, P., Teles, V., and Violette, S. (2005), *Dealing with spatial heterogenity*, Hydrogeology Journal, 13, 1, 161-183.

De Vries, L., Carrera, J., Falivene, O., Gratacos, O., and Slooten, L. (2009), *Application of Multiple Point Geostatistics to Non-Stationary Images*, Mathematical Geosciences, 41, 1, 29-42.

Delhomme, J.-P. (1979), *Spatial variability and uncertainty in groundwater flow parameters*, Water Resour. Res., 15, 281-290.

Feyen, L., and Caers, J. (2006), *Quantifying geological uncertainty for flow and transport modelling in multi-modal heterogeneous formations*, Advances in Water Resources 29, 6, 912-929.

Freeze, R. A. (1975), *A stochastic-conceptual analysis of one dimensional groundwater flow in non uniform homogemeous media*, Water Resour. Res., 11, 725-741.

Gelhar, L. W., Bakr, A. A., Gutjahr, A. L., and MacMillan, J. R. (1977), *Stochastic-conceptual analysis of 1-dimensional groundwater flow in nonuniform homogeneous media - Comment*, Water Resour. Res., 13, 2, 477-479.

Gelhar, L. W. (1986), *Stochastic subsurface hydrology from theory to applications*, Water Resour. Res., 22, 135S-145S.

Gómez-Hernández, J. J., and Wen, X.-H. (1998), *To be or not to be multi-gaussian? A reflection on stochastic hydrogeology*, Advances in Water Resources, 21, 1, 47-61.

Guardiano, F., and Srivastava, M. (1993), *Multivariate geostatistics: Beyond bivariate moments*, in *Geostatistics-Troia*, pp. 133-144, Kluwier Academic.

Hoffman, B. T., and Caers, J. (2007), *History matching by jointly perturbing local facies proportions and their spatial distribution: Application to a North Sea reservoir*, Journal of Petroleum Science and Engineering, 57, 3-4, 257-272.

Hu, L., and Chugunova, T. (2008), *Multiple-Point Geostatistics for Modeling Subsurface Heterogeneity: a Comprehensive Review*, Water Resour. Res., 44, W11413.

Huysmans, M., and Dassargues, A. (2008), *Application of multiple-point geostatistics on modelling groundwater flow and transport in a cross-bedded aquifer*, paper presented at GeoENV2008: 7th Int. Conference on Geostatistics for Environmental Applications, Southampton, 8-10 september 2008.

Journel, A., and Isaaks, E. (1984), *Conditional indicator simulation: Application to a Saskatchewan deposit*, Mathematical Geology, 16, 7, 685–718.

Journel, A., and Alabert, F. (1990), *New Method for Reservoir Mapping*, Journal of Petroleum Technology, 42, SPE paper 20781, 212-218.

Journel, A. (2002), *Combining Knowledge From Diverse Sources: An Alternative to Traditional Data Independence Hypotheses*, Mathematical Geology, 34, 5, 573-596.

Journel, A., and Zhang, T. (2006), *The necessity of a multiple-point prior model*, Math Geol, 38, 5, 591-610.

Journel, A. G., and Deutsch, C. V. (1993), *Entropy and spatial disorder*, Mathematical Geology, 25, 3, 329-355.

Kerrou, J., Renard, P., Hendricks-Franssen, H.-J., and Lunati, I. (2008), *Issues in characterizing heterogeneity and connectivity in non-multi-Gaussian media*, Advances in Water Resources, 31, 1, 147-159.

Koltermann, C., and Gorelick, S. (1996), *Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches*, Water Resour. Res., 32, 9, 2617-2658.

Langsholt, E., Kitterød, N., and Gottschalk, L. (1998), *Development of Three-Dimensional Hydrostratigraphical Architecture of the Unsaturated Zone Based on Soft and Hard Data*, Ground Water, 36, 1, 104-111.

Liu, Y., Harding, A., Abriel, W., and Strebelle, S. (2004), *Multiple-point simulation integrating wells, three-dimensional seismic data, and geology*, AAPG Bulletin, 88, 7, 905-921.

Mariethoz, G., Renard, P., Cornaton, F., and Jaquet, O. (2009), *Truncated plurigaussian simulations to characterize aquifer heterogeneity*, Ground Water, 47, 1, 13-24.

Mariethoz, G., Renard, P., and Froidevaux, R. (In Press), *Integrating collocated auxiliary parameters in geostatistical simulations using joint probability distributions and probability aggregation*, Water Resour. Res., doi:10.1029/2008WR007408.

Matheron, G. (1966), *Structure et composition des perméabilités*, Revue de l'IFP, 21, 564-580.

Matheron, G. (1967), *Eléments pour une théorie des milieux poreux*, Masson.

Neuman, S. P. (1984), *Role of geostatistics in subsurface hydrology*, in *Geostatistics for Natural Resources Characterization. Part 2*, pp. 787-816, D. Reidel Publishing Company.

Ortiz, J. M., and Deutsch, C. V. (2004), *Indicator Simulation Accounting for Multiple-Point Statistics*, Mathematical Geology, 36, 5, 545-565.

Pickard, D. (1980), *Unilateral Markov fields*, Advances in Applied Probability, 12, 655-671.

Renard, P. (2007), *Stochastic hydrogeology: what professionals really need?*, Ground Water, 45, 5, 531-541.

Ronayne, M., Gorelick, S., and Caers, J. (2008), *Identifying discrete geologic structures that produce anomalous hydraulic response: An inverse modeling approach*, Water Resour. Res., 44, W08426.

Sánchez-Vila, X., Carrera, J., and Girardi, J. P. (1996), *Scale effects in transmissivity*, Journal of Hydrology, 183, 1-2, 1-22.

Shannon, C. E. (1948), *A mathematical theory of communication*, The Bell system technical journal, 27, 379-423.

Stien, M., Abrahmsen, P., Hauge, R., and Kolbjørnsen, O. (2007), *Modification of the Snesim Algorithm*, paper presented at Petroleum Geostatistics 2007, 10-14 September 2007, EAGE, Cascais, Portugal.

Straubhaar, J., Walgenwitz, A., Renard, P., and Froidevaux, R. (2008), *Optimization issues in 3D multipoint statistics simulation*, paper presented at Geostats 2008, 1-5 Dec. 2008, Santiago, Chile.

Strebelle, S. (2002), *Conditional simulation of complex geological structures using multiple point statistics.*, Mathematical Geology, 34, 1, 1-22.

Strebelle, S., Payrazyan, K., and Caers, J. (2003), *Modeling of a deepwater turbidite reservoir conditional to seismic data using principal component analysis and multiple-point geostatistics*, SPE Journal, 8, 3, 227-235.

Strebelle, S., and Remy, N. (2005), *Post-processing of Multiple-point Geostatistical Models to Improve Reproduction of Training Patterns*, in *Geostatistics Banff 2004*, pp. 979-988, Springer

Strebelle, S. (2007), Simulation of Petrophysical Property Trends within Facies Geobodies, in *Petroleum Geostatistics*, edited, EAGE, Cascais, Portugal.

Suzuki, S., and Strebelle, S. (2007), *Real-time Post-Processing Method to Enhance Multiple-Point Statistics Simulation*, paper presented at Petroleum Geostatistics 2007, 10-14 September 2007, EAGA, Cascais, Portugal.

Western, A., Blöschl, G., and Grayson, R. (2001), *Toward capturing hydrologically significant connectivity in spatial patterns*, Water resources journal, 37, 1, 83-97.

Wu, J., Boucher, A., and Zhang, T. (2008), *A SGeMS code for pattern simulation of continuous and categorical variables: FILTERSIM*, Computers & Geosciences, 34, 12, 1863-1876.

Zhang, T., Switzer, P., and Journel, A. (2006), *Filter-Based Classification of Training Image Patterns for Spatial Simulation*, Mathematical Geology, 38, 1, 63-80.

Zinn, B., and Harvey, C. (2003), *When good statistical models of aquifer heterogeneity go bad: A comparison of flow, dispersion, and mass transfer in connected and multivariate Gaussian hydraulic conductivity fields*, Water Resour. Res., 39, 3, WR001146.

# Chapter 3

# Reconstruction of incomplete data sets or images using Direct Sampling*

*"What a man sees depends both upon what he looks at and also upon what his previous visual-conceptual experience has taught him to see."*

*Thomas S. Kuhn, The Structure of Scientific Revolutions*

---

* This chapter has been submitted for publication in Mathematical Geosciences as:
Mariethoz, G., P. Renard. Reconstruction of incomplete data sets or images using Direct Sampling.
The algorithms described in this article are protected by the international patent 2008WO-EP009819.

**Abstract** With increasingly sophisticated acquisition methods, the amount of data available for mapping physical parameters can be enormous. If the density of measurements is high enough, significant non-parametric spatial statistics can be derived from the data. In this context, we propose an adaptation of the geostatistical method of multiple-points by Direct Sampling (DS) for the reconstruction of partially informed images. The spatial patterns found in the data are mimicked without model inference. Therefore, no explicit assumptions are made on the spatial structure of the reconstructed fields, and the uncertainty on the spatial model can be neglected. The method gives good results for the reconstruction of complex 3D geometries from relatively small datasets. Moreover, very limited parameterization is needed.

## 1. Introduction

Missing or partial information on spatially distributed variables fields is a ubiquitous problem in Earth Sciences and is the major cause of (hydro)geological uncertainty. This state of matters is not likely to change, because there will always be an unsampled volume between wells or outcrops where no direct information is available. Geological heterogeneity further increases uncertainty because characterizing it asks for more data (Renard, et al., 2005).

On the other hand, numerical models such as finite elements or finite differences codes ask for exhaustive knowledge of the parameters of interest on the entire model domain. Unfortunately, the information content of geological surveys is often heterogeneous, with description of the geological features that can be very detailed at specific locations and completely absent elsewhere. The same problem arises with remote sensing or geophysical measurements, where uninformed zones may appear in an otherwise exhaustively informed domain. For example, geophysical surveys cannot access certain areas, producing shadow zones in the resulting parameters fields. Similarly, irregular monitoring or failure of recording devices can result in gaps in time series. Therefore we can only emphasize the need for robust methods able to complete sparsely known images and to deal with partial sampling. In this paper, we use the term of reconstruction to define this completion exercise.

Stochastic methods are the most advanced and appropriate way of dealing with spatial uncertainty (De Marsily, et al., 2005; Journel and Zhang, 2006). Among them, kriging has often been used in reconstruction problems, for example, for filling missing zones in geophysical surveys (Cornacchiulo and Bagtzoglou, 2004), reconstructing time series of aquifer piezometric level (Kumar and Ahmed, 2008) or for rain field reconstruction (Fiorucci, et al., 2001).

Stochastic simulations were also found to be adequate in a variety of cases. For geological applications, these methods are preferred when it is capital to preserve the heterogeneity observed in the measurements, which is the case for flow and transport simulations (Eaton, 2006). Several research papers develop such applications of geostatistical simulations (Guadagnini, et al., 2004; dell'Arciprete, et al., 2008). Other applications include, for example, the reconstruction of rock fracture surface (Marache, et al., 2002) or the generation of 3D permeability blocks from 2D thin sections in order to evaluate the permeability of samples (Youngseuk, et al., 2004).

The above mentioned methods proceed by first adjusting a parametric spatial model on the available data (typically one or a set of variograms), and then filling the unknown zones by properties honoring both the spatial model and the available measurements. With parametric techniques, adjusting the model can be tedious for complex 3D cases (Journel and Zhang, 2006). In this paper, we present a different approach based on the method of multiple-points simulations by Direct Sampling (DS) (Mariethoz, et al., submitted). By using non-parametric high-order statistics, this method is able to reproduce complex spatial structures that are present in the data, and therefore it has a high potential for addressing the issue of completing partial images. DS, like other multiple-points (MP) simulation techniques (Guardiano and Srivastava, 1993; Strebelle, 2002; Zhang, et al., 2006; Arpat and Caers, 2007; Straubhaar, et al., 2008), allows integrating an empirical and non-parametric structural concept in stochastic models. This non-parametric model takes the shape of a training image (TI) that depicts the spatial features of the variable of interest. The principle of MP simulations is to recompose the patterns found in the TI in order to obtain a field respecting various constraints, such as conditioning data or correlation with an auxiliary variable.

An important practical limitation of MP techniques is that finding the appropriate TI is not straightforward. Therefore it is common practice to build a TI using other geostatistical methods, such as Boolean simulations (Lantuéjoul, 2002), Sequential Indicator Simulations (Journel and Isaaks, 1984; Goovaerts, 1997), or a mix of different techniques (Journel and Zhang, 2006). This practice, although justified, leads to using multiple-points methods as a mere way of conditioning simulations coming from other methods. Hence, MP techniques do not take full profit of some of their most powerful advantages. Most of the possibilities offered by MP to integrate a non-parametric spatial model are hindered because the high order multiple-points model is fed by parametric, low order statistics.

Reconstructions using multiple-points statistics is possible and has been accomplished by Wu et al.(2008) using the FILTERSIM algorithm (Zhang, et al., 2006) or, by Okabe and Blunt (2004) to generate small scale 3D images of porous media. Nevertheless, inference of multiple-points statistics was done using TIs obtained by other methods, with all the above mentioned shortcomings. We propose a strategy for the inference of MP statistics based only on the available data. The idea is to abandon the use of a TI when the available dataset is large enough, and to extract MP statistics directly from the dataset, which becomes then a Training Dataset (TD). Indeed, the model information contained by the data is based on the true field, which is often more reliable than a training image derived from interpretation (provided that there are not large errors on the data), as indicated by Ortiz and Deutsch (2004). Because the model choice depends on the available data only, there is no room for model uncertainty (Kitanidis, 1986), even if uncertainty remains on the data measurements and their statistical significance.

Unlike traditional forms of MP simulation techniques that require a densely and regularly informed domain for inferring multiple-points statistics, DS can accommodate any data disposition and various kinds of variables (categorical and continuous). For most reconstruction problems, the available data, even if it is sparse and presents unknown zones, can be considered a TD from which high-order statistics can be derived. It opens the door for basing reconstructions on virtually any kind of data, and makes the technique appealing for a whole range of applications. Various cases of large TDs can be considered, with different levels of information to rely on for the reconstruction process. One extreme case consists of one unknown zone in an otherwise complete image. In this case, the TD contains information on spatial structures occurring at all scales. In the other extreme, the TD is sparse and spread on the entire domain.

Our method for reconstructing partial images uses the properties of the informed nodes to infer the values of the non-informed ones. An important underlying hypothesis is that these properties are stationary, i.e. that the statistical laws of the variable of interest are constant throughout the entire image (Koltermann and Gorelick, 1996). However, examples show that this hypothesis can be somewhat relaxed if enough information on non-stationarity is contained in the data.

The first part of the paper presents a summary of the DS algorithm, which is the main tool for solving reconstruction problems, and how it is used for the reconstruction of partial datasets. The rest of the paper is devoted to the presentation of examples, because we think it is the best way to grasp the principle, the practical advantages and limitations of this reconstruction method. Five reconstruction

examples are presented. The first example uses a simple 2D image to illustrate the difference in the result's quality when the TD is formed of continuous zones or is scattered through the entire domain. The second example reconstructs satellite photography of the Oahu River (New Zealand) in both categorical and continuous variables. The third example validates the method by performing a reconstruction based on a TD extracted from a 3D synthetic reservoir, therefore allowing comparison between the synthetic reference and the reconstruction. The fourth reconstruction example is based on a real 3D dataset made of a series of profiles digitized on a quarry front. The last example shows the application of the reconstruction method to borehole imaging data.

## 2. The reconstruction of partial images using Direct Sampling

Geostatistical simulation algorithms are aimed at producing realizations of a spatially correlated variable $Z$ at all $N$ locations $\mathbf{x}_i$ of a regular grid, with $i=[1,\cdots,N]$. Each such realization is a sample of the $N$-dimensional joint conditional cumulative distribution function (ccdf)

$$\mathrm{F}(\mathrm{z},\mathbf{x}) = \mathrm{Prob}\left\{Z(\mathbf{x}_1) \le z,\ Z(\mathbf{x}_2) \le z,\ \cdots,\ Z(\mathbf{x}_N) \le z\right\}. \quad (1)$$

Sequential simulation algorithms (Deutsch and Journel, 1992) are a practical way of sampling (1) by performing the following decomposition

$$\mathrm{F}(\mathrm{z},\mathbf{x}) = \mathrm{Prob}\left\{Z(\mathbf{x}_1) \le z\right\} \cdot \mathrm{Prob}\left\{Z(\mathbf{x}_2) \le z \mid z(\mathbf{x}_1)\right\} \cdot \ldots \cdot \mathrm{Prob}\left\{Z(\mathbf{x}_N) \le z \mid z(\mathbf{x}_1), \ldots, z(\mathbf{x}_{N-1})\right\} \quad (2)$$

.

The ccdf of $Z$ at each location depends of the ensemble of all previously determined $Z$ locations. Sequential simulations usually proceed by considering only a limited neighborhood (i.e. a subset of locations) of size $n$ for determining local ccdfs, with $n<<N$ to limit computational burden. At each location $\mathbf{x}$, the lag vectors $\mathbf{L}=\{\mathbf{h}_1,\ldots,\mathbf{h}_n\}=\{\mathbf{x}_1-\mathbf{x},..,\mathbf{x}_n-\mathbf{x}\}$ define the position of the informed neighboring locations. The data event at location $\mathbf{x}$ is defined by $\mathbf{d}_n(\mathbf{x},\mathbf{L})=\{Z(\mathbf{x}+\mathbf{h}_1),...,Z(\mathbf{x}+\mathbf{h}_n)\}$, the combination of the lag vectors and the value at the neighboring locations. Given this data event, sequential simulation techniques estimate at each successive location the ccdf for the variable of interest $Z$:

$$F(z,\mathbf{x},\mathbf{d}_n) = \text{Prob}\{Z(\mathbf{x}) \leq z \,|\mathbf{d}_n(\mathbf{x},\mathbf{L})\}. \quad (3)$$

Once the ccdf (3) is determined, a value for $Z(\mathbf{x})$ is drawn from it, and is thereafter considered as conditioning data when simulating the remaining nodes.

The DS algorithm (Mariethoz, et al., submitted) is a geostatistical sequential simulation algorithm making use of MP statistics. MP simulations derive the conditional distributions (3) from a training image (TI) representing the desired spatial structure of the variable of interest. The particularity of the DS lies in the fact that it does not store the probabilities associated to all pixels configurations found in the TI, whereas other algorithms store them either in a tree structure (Strebelle, 2002) or as lists (Straubhaar, et al., 2008). Instead, at each simulated node $\mathbf{x}$, the TI is sampled until the appropriate pattern is found. For each of the successive samples, the distance or mismatch $d\{\mathbf{d}_n(\mathbf{x},\mathbf{L}),\mathbf{d}_n(\mathbf{y},\mathbf{L})\}$ between the data event observed in the simulation $\mathbf{d}_n(\mathbf{x},\mathbf{L})$ and the one sampled from the training image $\mathbf{d}_n(\mathbf{y},\mathbf{L})$ is calculated ($\mathbf{y}$ denotes the nodes of the TI). If there is no mismatch or if it is lower than a given threshold, the sampling process is stopped and the value at the central node of the data event in the TI $Z(\mathbf{y})$ is directly pasted in the Simulation Grid (SG) at the location $\mathbf{x}$. At no point is the ccdf (3) needed.

Because the multiple-points statistics of the TI are not stored, there is no limit on the size of the neighborhood that can be used. Moreover, the neighboring nodes do not need to be located at fixed positions relatively to the central node (this set of positions, usually referred to as *template*, is necessary for traditional MP methods). The data event can have any geometry and can change for each simulated node. Therefore, it is possible to define the neighborhood as the $n$ closest previously simulated or conditioning nodes within a search radius $l$, where $n$ and $l$ are neighborhood parameters. This flexibility is a major advantage when neighborhoods are sparse or incomplete, which is precisely the case in the problem of reconstructing partial images.

Using different measures of distance also offers a high degree of flexibility. It is therefore possible to accommodate categorical as well as continuous variables. For categorical variables, the distance between a data event found in the simulation and another one found in the TI $d\{\mathbf{d}_n(\mathbf{x},\mathbf{L}),\mathbf{d}_n(\mathbf{y},\mathbf{L})\}$ is given by the proportion of non-matching nodes. It is calculated by using the indicator variable $a$ that equals 0 if two nodes have identical value and 1 otherwise:

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}=\frac{1}{n}\sum_{i=1}^{n}a_i \quad \in [0,1], \quad where\ a_i=\begin{cases}0 & if\ Z(\mathbf{x}_i)=Z(\mathbf{y}_i)\\ 1 & if\ Z(\mathbf{x}_i)\neq Z(\mathbf{y}_i)\end{cases}. \tag{4}$$

A measure of distance able to accommodate continuous variables is the normalized pair wise Manhattan distance,

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|Z(\mathbf{x}_i)-Z(\mathbf{y}_i)\right|}{\mathrm{argmax}(Z^{ti})-\mathrm{argmin}(Z^{ti})} \quad \in [0,1]\cdot \tag{5}$$

The normalizing factor is the largest interval within the $Z$ values found in the TI, and is used to ensure that the distance values are comprised in the interval [0,1]. Without this normalization, the choice of a threshold value would be made more difficult. For example, numerical tests have shown that 0.05 is a low threshold and 0.5 a high threshold. Such guidelines can be more difficult to establish without normalization.

Simulations resulting from DS reproduce the spatial properties of the scanned image. Even complex properties are reproduced because the method allows exploiting high order statistics from a data set. Contrary to other MP methods, DS does not need multiple-grids (Strebelle, 2002) for reproducing structures at different scales. Instead, this is accomplished by accommodating data events of different sizes. As such, it does not encounter problems of relocation of conditioning data on the different simulation grids (Saripally, 2008), making it easy to integrate large data sets. As the reconstruction problem relies on large datasets, this feature is a major advantage.

The core of the DS algorithm remains unchanged if a dataset has to be scanned instead of a TI. This allows simulating unknown nodes by scanning the informed ones only. The goal is that the unknown zones of the SG are filled with patterns coming from the TD. If a large diversity of patterns is present in the TD, this diversity can be reproduced in the unknown zones of the SG. The DS algorithm as described in Mariethoz, et al. (submitted) has to be slightly modified in order to scan the TD instead of a TI. The major change in the DS algorithm is that **y** does not designate any more a location in the TI, but a location in the TD once it has been migrated on the SG. The resulting algorithm is summarized in the following steps:

1) All points of the TD are assigned to the closest grid nodes in the SG (for more clarity, the ensemble of the SG nodes where TD points have been migrated is designated TD hereinafter). When several conditioning data should be assigned

to the same grid node, we assign the closest one to the grid node. Other alternatives are to assign the most frequent one, or to sample randomly one of these values for each simulation.

2) Define a path through the remaining nodes of the SG.
3) For each successive location **x** in the path:
   a) Find the neighbors of **x**. They consist of a maximum of the *n* closest grid nodes $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ that were already assigned or simulated in the SG and located within a search radius *l*. Compute the lag vectors $\mathbf{L} = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\} = \{\mathbf{x}_1 - \mathbf{x}, .., \mathbf{x}_n - \mathbf{x}\}$ defining the neighborhood of **x**.
   b) Define the data event $\mathbf{d}_n(\mathbf{x}, \mathbf{L}) = \{Z(\mathbf{x} + \mathbf{h}_1), ..., Z(\mathbf{x} + \mathbf{h}_n)\}$. It is a vector containing the values of the variable of interest at all the nodes of the neighborhood.
   c) Randomly draw a location **y** in the TD and from that location scan systematically the TD. For each location **y**:
      i) Find the data event $\mathbf{d}_n(\mathbf{y}, \mathbf{L})$.
      ii) Compute the distance $d\{\mathbf{d}_n(\mathbf{x}, \mathbf{L}), \mathbf{d}_n(\mathbf{y}, \mathbf{L})\}$ between both data events. If the lag vectors point to nodes of the SG that are not informed by the TD or that are outside the SG, these nodes are counted as false ($a_i = 1$) if the variable to reconstruct is categorical (distance (4)). In case of continuous variable (distance (5)), they are given the maximum possible distance, which is the largest interval within the *Z* values found in the TD.
      iii) If it is the lowest distance obtained so far in the scanning of the TD, store **y,** *Z*(**y**) and $d\{\mathbf{d}_n(\mathbf{x}, \mathbf{L}), \mathbf{d}_n(\mathbf{y}, \mathbf{L})\}$.
      iv) If $d\{\mathbf{d}_n(\mathbf{x}, \mathbf{L}), \mathbf{d}_n(\mathbf{y}, \mathbf{L})\}$ is smaller than the acceptance threshold *t*, the value *Z*(**y**) is assigned to *Z*(**x**). The algorithm proceeds then to step 3 in order to simulate another node of the SG.
      v) If no satisfying distance under the threshold has been found in the TD, the node **y** with the lowest distance is accepted and its value *Z*(**y**) is assigned to *Z*(**x**).

The data event $\mathbf{d}_n(\mathbf{x}, \mathbf{L})$, including its shape given by the lag vectors, may not correspond exactly to any informed pattern in the TD. In those cases, the distance measure favors the selection of a data event in the TD that maximizes the number of

correctly informed nodes. This is done by attributing the maximum distance to the non-informed nodes in the data event (point 3.c). This is equivalent to reducing the order of the statistics when no suitable data event is found in the TD. In the worst case, there will be no location in the TD corresponding to the lags vectors searched for, and the marginal probabilities will then be sampled.

There are only a limited number of parameters governing the algorithm. The main ones are the maximum number of neighbors $n$ and the threshold $t$. Setting $n$ to a large value (between 30 and 50 nodes) and $t$ to a very low value (typically 0) guarantees to yield the best possible results DS can provide, sometimes at a high cost in terms of CPU time. Therefore, we used $t = 0$ for all the examples presented below, except the continuous variable case (because a perfect match can never be found between continuous variable data events, and it makes no sense to set $t = 0$ in this case).

## 3. Reconstruction examples

### 3.1. Spatial repartition of the TD

The first reconstruction example illustrates how the spatial scattering of the TD influences the reconstruction results. Figure 1a shows a categorical image, representing sand channels in a clay matrix (Strebelle, 2002). The central zone (36% of the entire image) is considered unknown and it is reconstructed by scanning the edges of the image with a neighborhood constituted of the 40 closest nodes. Three realizations of reconstruction are shown in Figure 1b to Figure 1d, generated using the algorithm described above. Distance (4) was used in the DS simulation because the variable of interest is categorical.

The resulting realizations have the features one would expect when looking at the original training image. This reconstruction is possible because the training data set contains information at all scales, such as many small portions of channels on the east and west sides of the image and large connected features in the North and South. This example is based on a dataset that is incomplete, but that contains spatial information on a variety of scales. The presence of data on large and continuous zones results in the presence of small and large scale patterns. This diversity of patterns present in the TD extensively feeds the MP model. Nevertheless, several realistic situations may be different. There are many cases where adjacent grid nodes are not informed. With dispersed data, small scale patterns have less chance to occur

in the TD, tending to produce a degenerated MP model. Moreover, the absence of continuous conditioning zones leads to a lack of information at small scale.

Figure 2 shows the results of reconstruction using three different sparse TDs coming from the image of Figure 1a. The first TD consists of 5,000 nodes taken at random locations (8% of a total of 62,500 nodes in the image grid). It is displayed in Figure 2a, with the corresponding reconstruction in Figure 2b. Figure 2c and Figure 2e show larger TDs of respectively 10,000 and 20,000 informed nodes (16% and 31%), and their corresponding reconstructions in Figure 2d and Figure 2f.

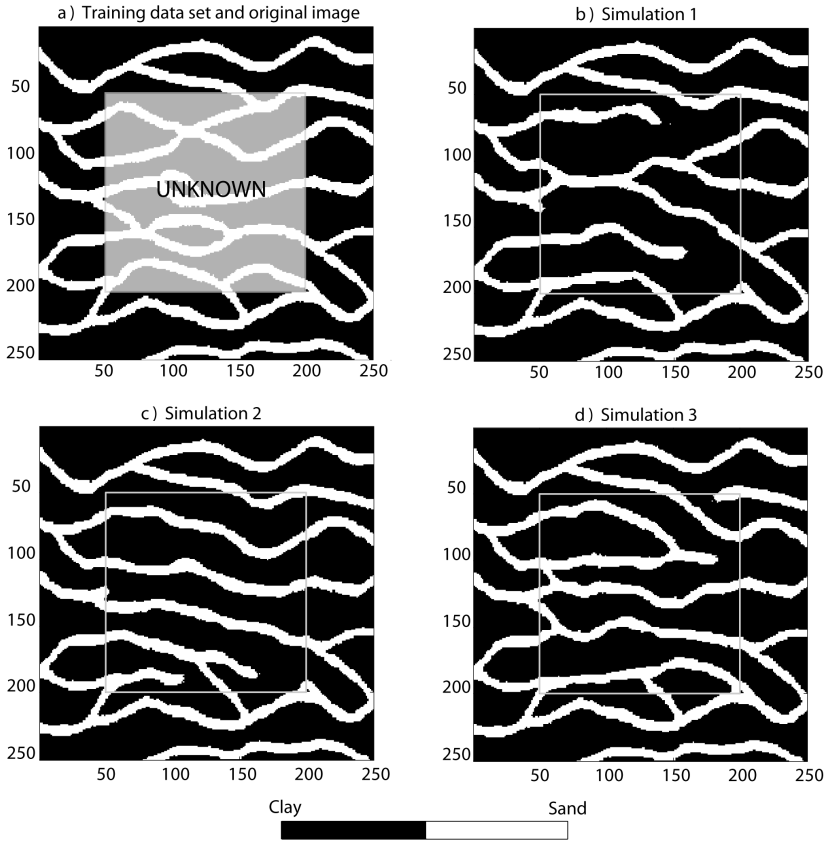



**Figure 1** The reconstruction of a simple categorical image with a hole (36% of the entire image). a) The original complete image (Strebelle, 2002) and the unknown zone to reconstruct. b), c) and d) Three realizations of reconstruction of the central zone.

The random position of the informed nodes explains the relatively low quality of the reconstructed images for the TD of 5,000 informed nodes (Figure 2b), where small scale patterns are virtually absent (see zoomed part of Figure 2a). Gómez-Hernández and Wen (1998) show that using point data only, it is impossible to know whether the high values of hydraulic conductivity are connected or not. The reconstruction method does not make any assumption on the spatial continuity of the variable to simulate. If this spatial continuity is not represented in the TD, it will also not be present in the reconstruction. Hence, the resulting reconstructions are greatly affected by the lack of information at small scale. Conversely, Figure 2d and Figure 2f show that an important fraction of informed nodes contains enough information to reconstruct accurately the entire image, because such a dense dataset includes both large and small patterns (zoomed parts of Figure 2c and Figure 2e).

To emphasize the lower performance of DS for reconstruction based on scattered datasets, it has been compared to Sequential Indicator Simulation (SIS, Deutsch and Journel, 1992) using the same 5,000 conditioning data as Figure 2b. The variogram model used for SIS was adjusted on the 5,000 data TD. It is a spherical model, with ranges of 27 m in the EW direction and 10 m in the NS direction. The same 5'000 data were used for conditioning. The results of this reconstruction using SIS are shown in Figure 3. It displays better connectivity of sand bodies than DS reconstruction, even if artifacts appear due to the anisotropic variogram structure.

This comparison shows that DS reconstruction should better be used with TDs that are large and that contain patterns of different scales. In order to capture all scales, an appropriate sampling strategy would consist of densely measured regions and sparse measurements in other areas (this is valid for all simulation/estimation methods). This said, such a limitation does not exclude all practical applications of the reconstruction method. For geological applications, outcrops and quarry fronts provide dense and multi-scale information on 3-dimensional geological structures. To a lesser extent, well logs also provide small scale patterns in the vertical direction. Remote sensing applications are also numerous, often involving problems similar to the setting of Figure 1a, where one well-defined zone is unknown in an otherwise complete image.

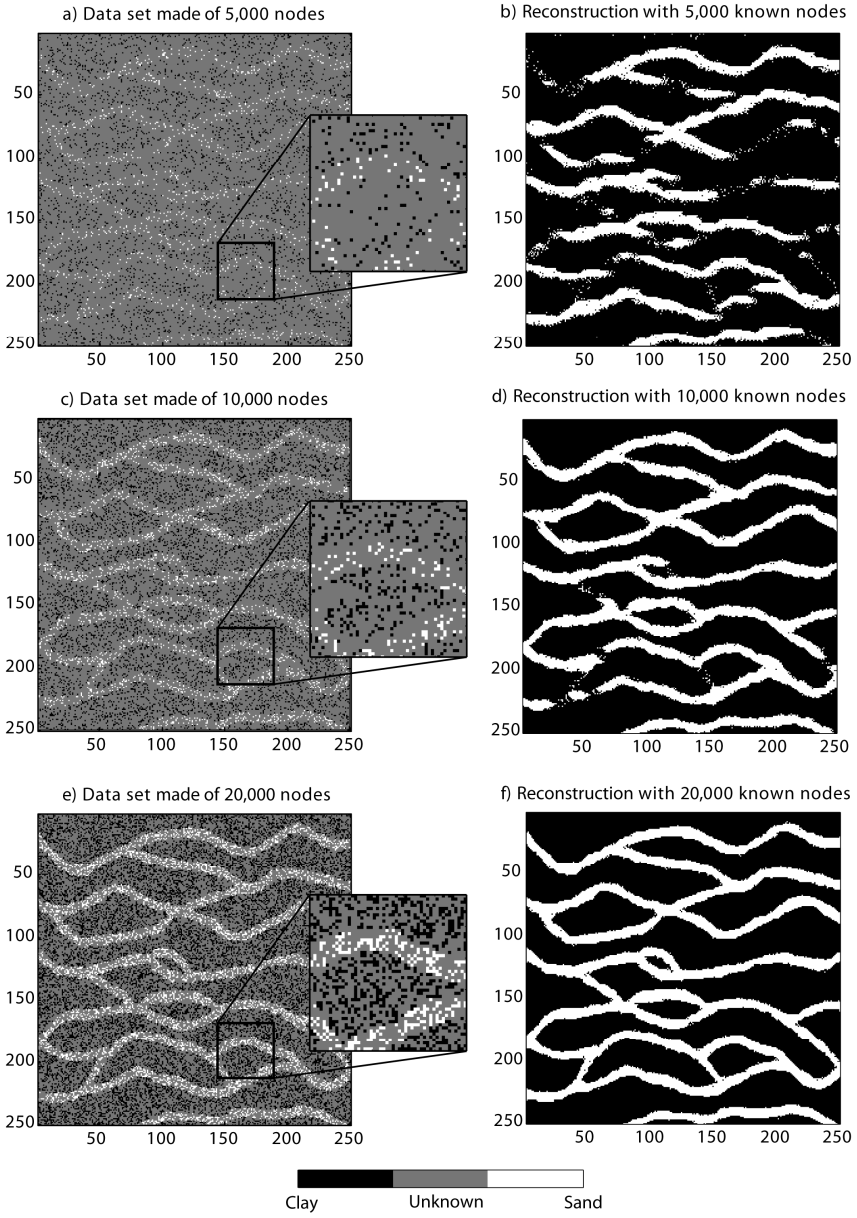



**Figure 2** Reconstruction using scattered datasets. a) Dataset consisting of 5000 nodes. b) One realization of reconstruction with a dataset of 5,000 nodes

(8% of the entire image). c) and d) same with 10,000 informed nodes (16%). e) and f) same with 20'000 informed nodes (31%).

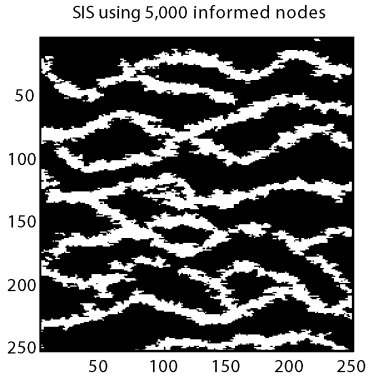



**Figure 3** Reconstruction using SIS with the same dataset of 5,000 nodes.

## 3.2. Continuous variable example

The second example illustrates the same problem using the synthetic (but realistic) transmissivity field presented by Kerrou et al. (2008). The reference image was built from an aerial photograph of braided channels in the Ohau River, New Zealand (Mosley, 1982). The image size is 1000m by 400m and it has been discretized on a grid of 440 by 176 nodes. It is divided in two facies codes: 0 for gravelly channels and 1 for silty lenses. Facies were converted to a continuous transmissivity field as described by Kerrou et al. (2008). Two multiGaussian unconditional simulations were separately generated to populate the channels and the lenses with logarithm of transmissivity values. The first simulation (channels) uses an exponential variogram with correlation range of 3 m. The second simulation (lenses) has a nested variogram that includes one isotropic exponential model with a 3m range, plus a cubic anisotropic model with a range of 600 m in the EW direction and 300 m in the NS direction. This leads to a bimodal, non-multiGaussian, anisotropic transmissivity field.

Similarly to the previous example, Figure 4a shows the extension of the unknown zone, which represents 20% of the domain. Figure 4b displays one realization of reconstruction. Distance measure (5) was used since the variable is continuous. In this case, a neighborhood of only 20 nodes was enough to give satisfying results, with a threshold set at $t$ = 0.025. Visually, the reconstructed

central zone has a similar spatial structure to the rest of the image. Figure 4c displays the mean of 300 reconstruction simulations, which come close to a conditional estimation. It shows that there is a zone of low uncertainty on the edges of the unknown zone. The features in this zone of low uncertainty are coherent with the position of the channels and lenses in the reference field.

Figure 4d to Figure 4f illustrate the same reconstruction for the corresponding categorical variable, using the measure of distance (4) and the same neighborhood. Structures are visible in the central part of the estimation, showing that uncertainty is reduced even far from the edges.

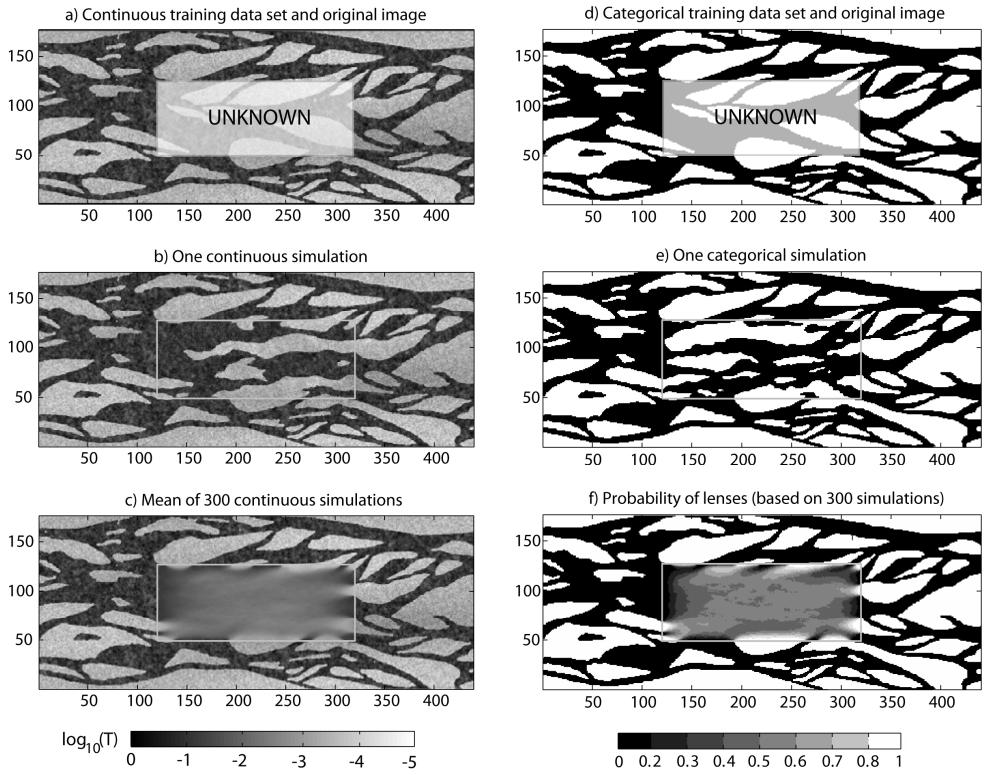



**Figure 4** Reconstruction of an image depicting an aerial view of a braided system. The size of the image is 440 by 176 nodes. a) The original complete image (Kerrou, et al., 2008) and the unknown zone to reconstruct. b) One realization of reconstruction of the central zone. c) Mean $\log_{10}T$, based on 300 realizations. d) to f) Same reconstruction for categorical variable. Here, f) is the probability of occurrence of lenses.

Two sorts of constraints are imposed on the reconstruction of the unknown zone. The first one is given by the patterns present in the TD and that DS reproduces in the reconstruction. This is the structural constraint. The second constraint is imposed by the conditioning nodes that are close to the unknown zone. This is the local constraint. Ideally, these two constraining factors should be compatible as they both come from a single TD. In practice, non-stationarity can be present and the local conditioning might become antagonistic with structural constraints. This situation occurs in the present example. The reference image has a global proportion of lenses of 0.50, but it is imbalanced between the different zones, with a proportion of lenses of 0.44 in the central (unknown) part and 0.51 in the outside. Figure 5a highlights the area of the TD within 15 pixels from the unknown zone, where the proportion of lenses is only 0.39, much lower than the marginal proportion. The result of this antagonism is a reconstruction that has statistical properties lying in between the ones of the entire TD and the area of the TD close to the unknown zone. Figure 5b shows the histogram of the proportion of lenses in 300 reconstructions. It is centered in between the proportions found in the TD and the proportions found in the 15 pixels "crust" around the unknown zone.

The same phenomenon happens for the reproduction of the histograms (Figure 5c) of the simulated $\log_{10}T$ values. Here again, most histograms of simulated values are halfway between the TD and the conditioning 15 pixels crust. The difference in lenses proportion that was noted for Figure 5b is visible in the second mode of the distributions. Similarly, one can observe quite different omnidirectional variograms for the TD and the 15 pixels crust (Figure 5d). The reconstructions should have a variogram similar to the one of the TD, because they are made of patterns of nodes coming directly from the TD. Nevertheless, coherence with the existing bodies of the external area generates a structure that tends to the one of the area of the TD neighboring the unknown zone. The resulting variogram lies in between these two models, which is necessary in order to have a coherent reconstruction.

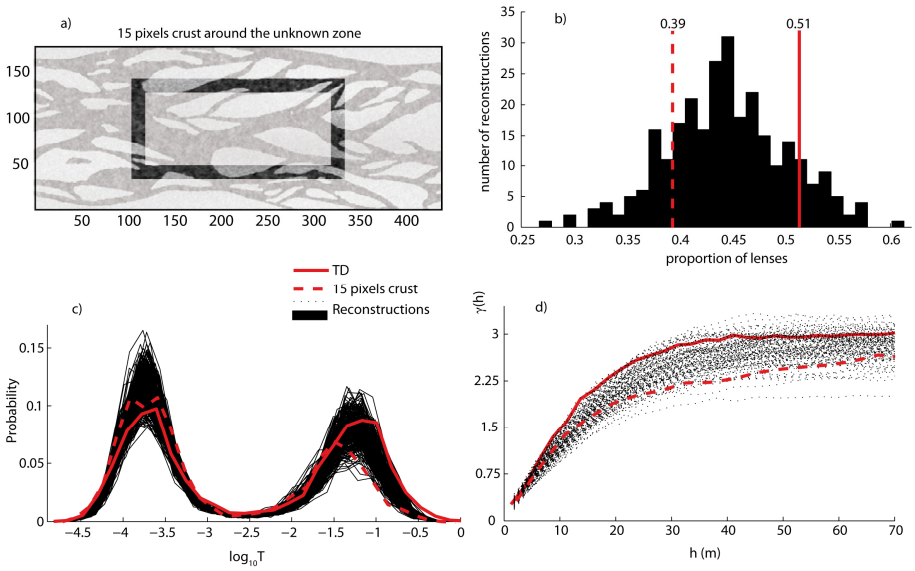



**Figure 5** Illustration of local and structural constraints. a) 15 pixels crust around the unknown zone. b) Proportion of lenses in 300 continuous variable reconstructions simulations and in the different areas of the reference image. c) Histograms of $\log_{10}T$ for 300 continuous variable reconstructions, for the 15 pixels crust and for the entire TD. d) Experimental omnidirectional variograms for 300 continuous variable reconstructions, for the 15 pixels crust and for the entire TD.

## 3.3. 3D synthetic example

The third example shows how much information can be extracted from a TD composed of continuous zones, such as 2D slices in a 3D bloc. The reference image is an object-based simulation depicting a geological setting of turbidites on a continental margin (Strebelle, et al., 2002). This reference image (Figure 6a) is non-stationary in both horizontal and vertical directions because its scale is very large, ranging from proximal turbiditic sediments to distal pelagic deposits in the abyssal plain. It is made of 6 geological facies and discretized in 70 x 183 x 20 grid nodes. From facies code 0 to 5, the different lithologies correspond respectively to: abyssopelagic sediments, bathypelagic sediments, epipelagic sediments, coarse turbiditic sediments, medium turbiditic sediments, fine turbiditic sediments.

The TD (Figure 6b) is created by extracting points along 5 vertical cross-sections in the reference image. In a real setting, these slices would typically correspond to

interpreted geophysical surveys, such as seismic profiles. The reconstruction grid has the same discretization as the reference image, with a total of 256,200 nodes, 11,380 of which (4.4%) being informed by the extracted slices.

For the reconstructions, a neighborhood of 20 nodes was again used. In order to ensure that the first simulated nodes are close to the informed locations (close to the slices), a custom path in the SG was used instead of a random path. It consists of an altered random path ensuring that each simulated node has at least one of its direct neighbors informed (direct neighbors are only one grid block away from the simulated node). Therefore, the nodes close to the informed cross-sections are simulated first, and the simulation proceeds away from there on.

If a random path would be used, it would start simulating nodes far away from the informed cross-sections. Because large patterns have less chances to be found in the TD, it is likely that low order (or even marginal) probabilities will be sampled at the beginning of the simulation. This would incur inconsistencies in the large patterns, which may deteriorate the resulting images.

20 realizations of reconstruction were performed, with one of them shown in Figure 6c. In average, 75.3% of the nodes were correctly simulated (i.e. they were attributed the same value as in the reference image). This proportion of correctly simulated nodes is remarkably constant through the 20 simulations, with a minimum of 73.1% and a maximum of 77.2%. Although the shape of the turbiditic lobes is somewhat degraded, the lobes structure is still visible with a surprisingly high accuracy given the low information content of the TD. The information on facies 3 is very scarce because it is hardly present in the TD, but the reconstruction algorithm could manage to produce elongated bodies similar to what is found in the reference. The hierarchy of the turbiditic structures is also correctly reconstructed, with the channels coming first, followed by the proximal and distal parts of the lobes. Figure 7a and Figure 7b show the probability maps of facies 4 and 5 (areas with a probability higher than 0.7 are highlighted), where one can observe that facies 4 is generally more distal than facies 5, and that facies 5 is surrounded by facies 4 on the distal side. Corresponding structures are observed in the reference image, but cannot be deduced easily by looking at the TD only.

The non-stationary trends are also correctly addressed by the reconstruction procedure. Information on non-stationarity is present in the TD because cross-sections are large enough to capture it. Figure 7c displays the probability map of facies 0 (probability higher than 0.7). This facies is only present in the abyssal zone, as it should according to the reference image. Similarly, turbiditic structures only appear on the proximal part of the image, along with facies 1 and 2.

On Figure 7c, a sharp transition in the probability map of facies 0 is visible at the equidistance of two parallel TD cross-sections (right side of the image). We believe such artifacts are related to the custom path used. The simulation grid is filled starting with the nodes close to the data. When simulated nodes originating from distant slices meet, incompatibilities can occur. One way to deal with these artifacts could be to use larger neighborhoods. This would increase the probability that the simulated nodes coming from another slice belong to the neighborhood of the node to simulate. A more proper solution would be to use syn-processing to remove these artifacts (Mariethoz, et al., submitted). It is a recursive procedure that recursively un-simulates and re-simulates nodes with the aim of removing inconsistencies in the patterns of the simulation.

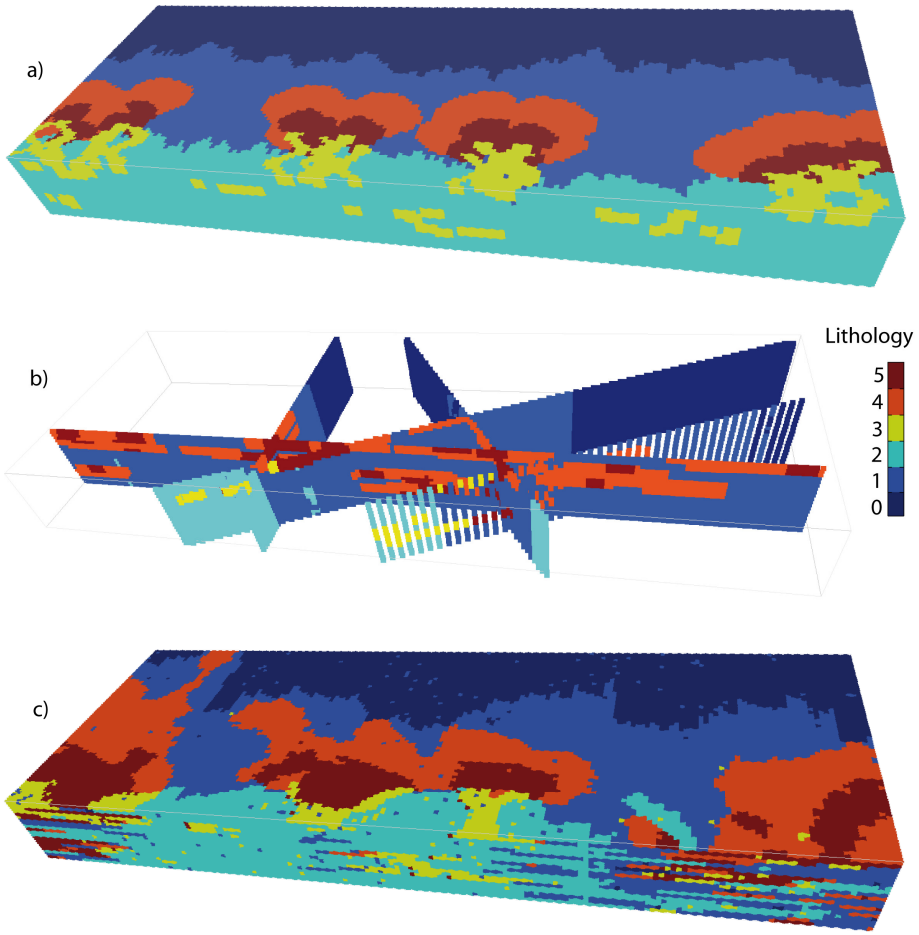



**Figure 6** Reconstruction of turbidite lobes. a) Original reference image, from which the dataset consisting of 5 cross-sections b) has been extracted. c) One realization of reconstruction.

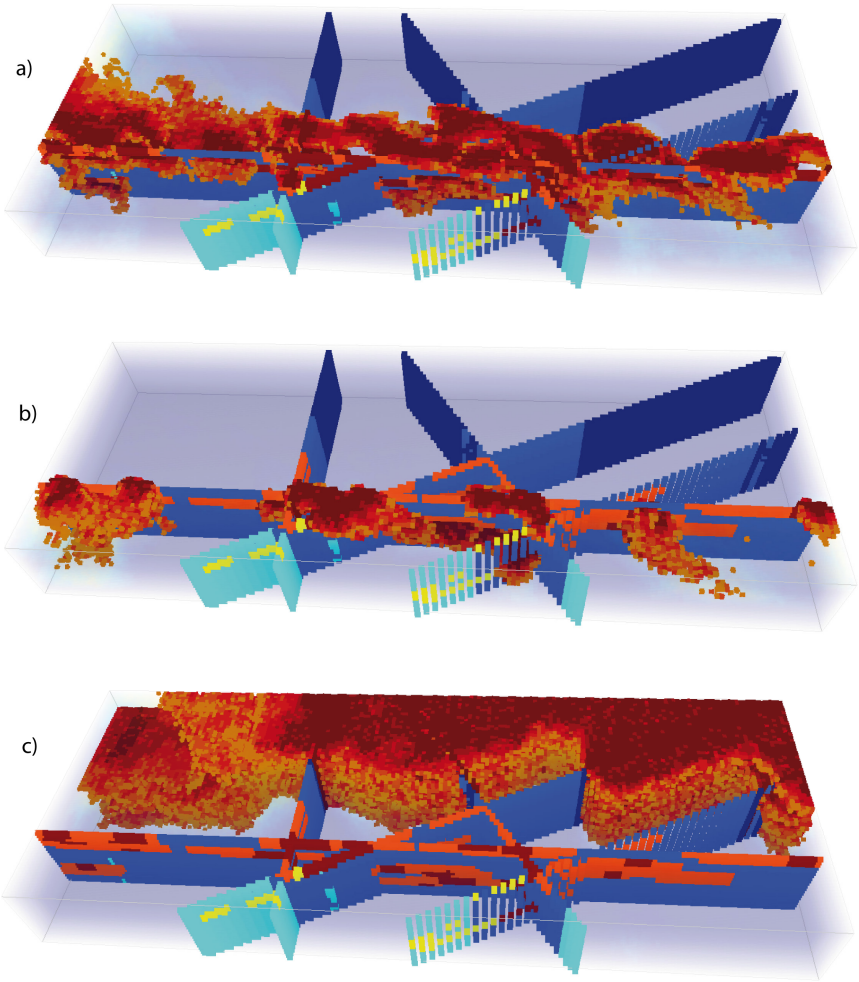



**Figure 7** Probability maps of facies, computed from 20 realizations. a) Area where the probability of facies 4 exceeds 0.7. b) Area where the probability of facies 5 exceeds 0.7. c) Area where the probability of facies 0 exceeds 0.7.

Figure 8 displays the facies proportions in the reference, in the simulations and in the TD. In the TD, there is a large bias for most of the facies because the irregular disposition of the slices and the global non-stationarity of the reference image result in a sampling that is not representative. Facies 0, 2 and 3 are underrepresented and

facies 1, 4 and 5 are overrepresented. In the reconstructions, the largest biases (facies 0, 1, 4 and 5) have been reduced. Structural constraints are not well defined because the TD is not fully representative and because of non-stationarity. On the other hand, local constraints take very well into account the non-stationarity inherent to the image. As observed for the previous example, even if the structural constraints are biased, the combination with the local constraints corrects the structure to a certain extent. The result is that the facies proportions in the reconstructions are closer to the reference proportions than what is observed in the TD. Here again, accounting for multiple-point statistics allows exploiting the large dataset with maximum coherence, even in the presence of non-stationarity.

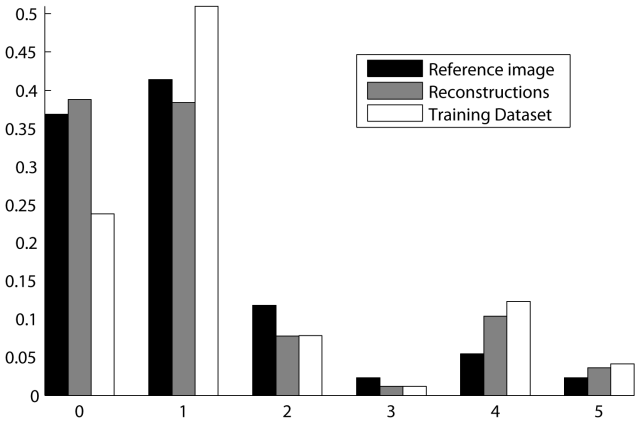



**Figure 8** Facies proportions in the reference image, the 20 reconstructions and the TD. Proportions of reconstructions have been computed using 20 realizations.

## 3.4. Real case 3D application

The next example is the application of the DS reconstruction algorithm to a real 3D dataset. The study site is a quarry located in the sediments of the Lambro River, in northern Italy. The geological environment is an extremely heterogeneous setting of meandering sediments with point-bar deposits. The geometries of such sedimentological formations are widely analyzed and described in the literature (Miall, 1996; Nichols, 1999; Kostic and Aigner, 2007) and are a classical challenge for geostatistical simulations (Deutsch and Tran, 2002; Zappa, et al., 2005; Rubin, et al., 2006; Mariethoz, et al., 2009).

The available data are based on a series of measurements taken on 5 different exposure stages of the quarry front. On each of the 5 documented fronts, centimeter-scale logging, textural and petrophysical analysis have been performed. Hierarchical classification and interpretation of the sediments have led to the definition of 4 hydrofacies, represented in a dataset of 483,472 point values located along the 5 quarry fronts (details can be found in Bersezio, et al., 2007). Migrated on the SG, these points constitute the Training Dataset used for the reconstructions (Figure 9a). A previous study (dell'Arciprete, et al., 2008) performed reconstructions based on the same dataset, using well-established methods of SIS (Isaaks, 1984) and transitions probabilities (Carle and Fogg, 1997). Contrary to dell'Arciprete, et al. (2008), the present application does not intend to benchmark methods, but only to apply DS reconstruction to a real dataset.

Two different grids are used for reconstruction. The first one, coarse, is constituted of 2,764,800 grid blocks. The second one, finer, has a total of 9,768,000 smaller blocks. Because of the very high spatial resolution of measurements, each facies point data cannot be migrated on a separate grid block, even on the fine grid. The number of informed grid nodes is different on the two grids, but in both cases it remains a very low proportion of the entire grid (1.31% for the coarse grid and 0.87% for the finer grid). A summary of the grid properties is presented in Table 1.

| | **Size of grid x,y,z** | | | **Total nb of nodes** | **Blocks size x,y,z (cm)** | | | **Nb of informed nodes** |
|---|---|---|---|---|---|---|---|---|
| **Coarse grid** | 160 | 240 | 72 | 2,764,800 | 31.25 | 31.25 | 12.5 | 36,291 (1.31%) |
| **Fine grid** | 240 | 370 | 110 | 9,768,000 | 20 | 20 | 8 | 85,056 (0.87%) |

**Table 1** Summary of the properties of both reconstruction grids.

On the finer grid, a single reconstruction realization was made to keep CPU time reasonable. Figure 9b and Figure 9c show two different views of the reconstructed lithologies. Figure 10 displays 6 cross-sections of the reconstruction on fine grid. 20 other realizations were performed on the coarse grid. These simulations are used to compute statistics on the probability of occurrence of facies (Figure 11). All reconstructions use a neighborhood of 35 nodes and the same custom path as the in previous example.

The purpose of the reconstruction on fine grid is to check the reproduction of the very fine stratified structures present in the TD, and a single realization is enough for this. Figure 9b displays the TD together with cross-sections in the SG. The intersections of all interfaces between the different transects are coherent. The stratifications observed in the TD (alternating beddings of lithologies 2, 3 and 4 are clearly visible on the reconstruction.

Figure 10 shows that the non-stationarity of the data is preserved. For example, there is more stratification of lithofacies 3 and 4 with increasing y coordinate (compare Figure 10a, Figure 10c and Figure 10f). Moreover, one can observe large bodies of lithofacies 3 in a matrix of facies 2 on Figure 10a, whereas smaller and less continuous bodies of facies 2 are found at the other end of the domain. Similarly, facies 1 is only present in certain areas of the TD. The reconstruction respects this spatial distribution as facies 1 is essentially found on Figure 10a and Figure 10b, and is almost absent of the other cross-sections.

Structures showing different orientations are present in the TD. This is especially visible with facies 4 that is presents in continuous sub-horizontal layers in the upper part of the domain, but also dipping structures having an angle of about 45° in the lower part of the domain (Figure 9a). Modeling such structural non-stationarity would normally call for dividing the domain in separate structural units that would be simulated separately. This has not been done here. Nevertheless, Figure 10c, Figure 10d and Figure 10e show that this dual orientation of structures is present in the reconstruction.

Figure 11 shows that the lithofacies proportions in the TD and in the reconstructions are very close, but there is no exact match. As the proportions of the real field are not known, it is not possible to know if the reconstructed proportions are closer to the reality than the proportions of the TD.

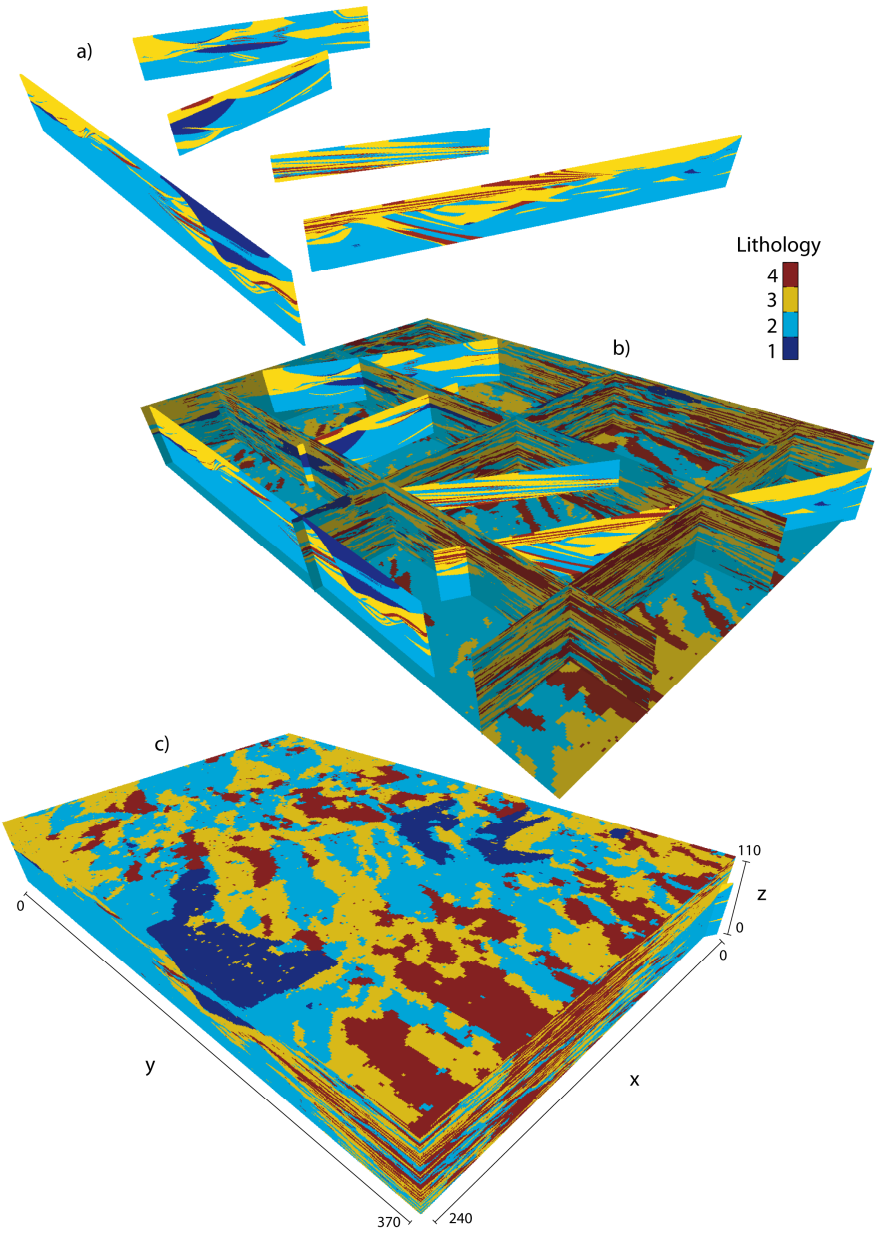



**Figure 9** a) Original dataset consisting of 5 interpreted cross-sections. b) One high-resolution simulation of reconstruction, slices view (10 millions cells). c) Same simulation of reconstruction, block view.

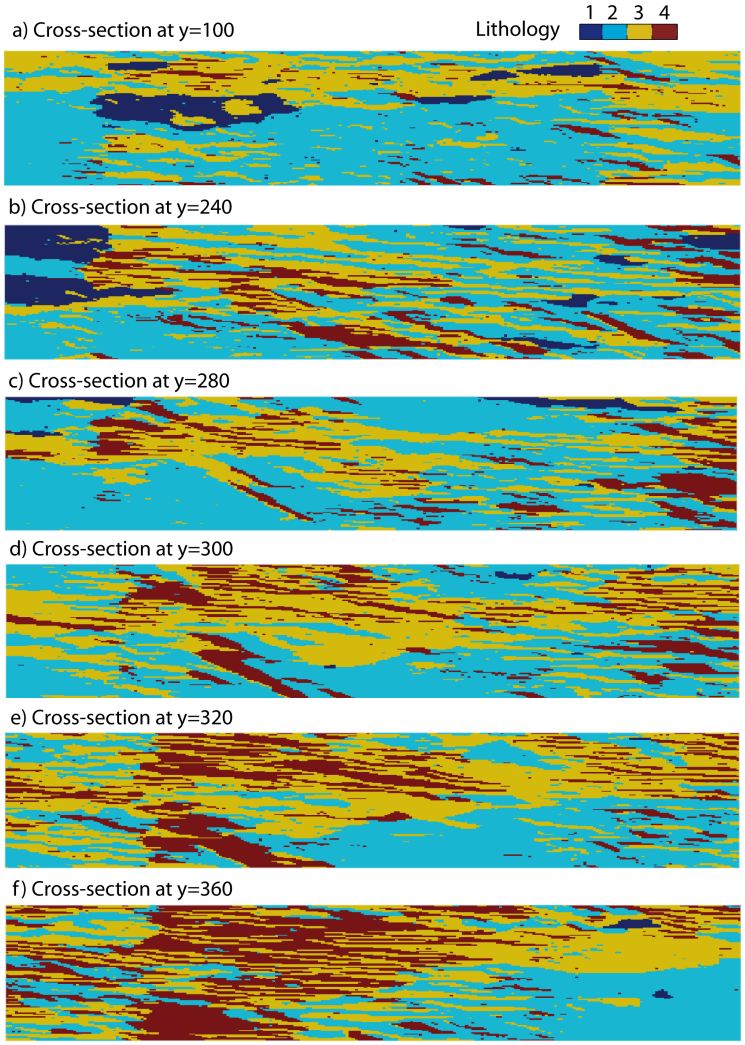



**Figure 10** a) to f) Cross-sections in the reconstruction on fine grid. All sections are parallel to the x axis, at a constant y coordinate indicated on the figure.

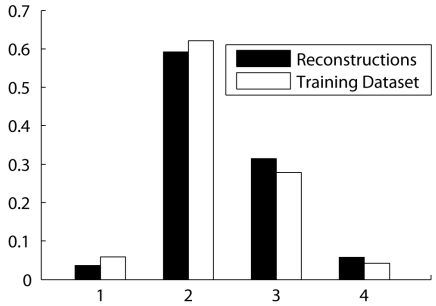



**Figure 11** Facies proportions in the TD and in 20 reconstruction realizations on the coarse grid.

## 3.5. Real case borehole imagery example

The last reconstruction example illustrates an application of the DS reconstruction method to a borehole microresistivity image. Borehole images are a powerful way of obtaining data for the orientation of bedding planes and fractures (Wu and Pollard, 2002; Khoshbakht, et al., 2009) to identify cycles in sedimentary sequences (Lefranc, et al., 2008), or for visual inspection of boreholes.

Borehole imagery is a common example of unknown zones in an otherwise known image. Borehole images can be obtained by inserting tools such as a FMI (Fullbore Formation MicroImager, Schlumberger) device into the borehole. The device is composed of 4 pads carrying 2 sensors each. The sensors induce a current into the formation that is modulated in amplitude, providing rich information about the petrophysical properties. The resulting images are exhaustively informed along the pads trajectory, but not informed at all between the pads. Therefore, accurate estimation of bedding planes and fractures orientation calls for interpolation and reconstruction techniques. Reconstruction could also be a precious aid for the visual inspection of the inside of a borehole. Moreover, it is much easier to carry on geological interpretation based on complete images.

Figure 12a displays a dataset obtained with a FMI device. The unknown zones caused by the spacing between pads are shown in white. In this case, 70% of the image is informed, constituting the TD available for reconstruction.

20 realizations of reconstruction have been generated using distance (5) with a neighborhood of $n$=20 nodes and a threshold of $t$=0.01. One of them is displayed in Figure 12b. The fine scale heterogeneity and the continuity of horizons and fractures are preserved. At locations where they are present in the TD, bedding planes are

connected (e.g. dark lines in the central part of the image). The internal texture within the sedimentary units (grey and white) is also ensured. The TD is non-stationary because it is made of different zones comprising thick and thin horizons, fractured and non-fractured areas, massive or finely laminated structures. Here again, the method is able to accommodate the non-stationarity because the different zones of the TD are not mixed in the reconstructed areas.

In Figure 12c, the mean of the entire stack shows that the continuity of the major features is well reproduced on all simulations, with more variability on the small scale heterogeneity resulting in blurry areas along the unknown zones.

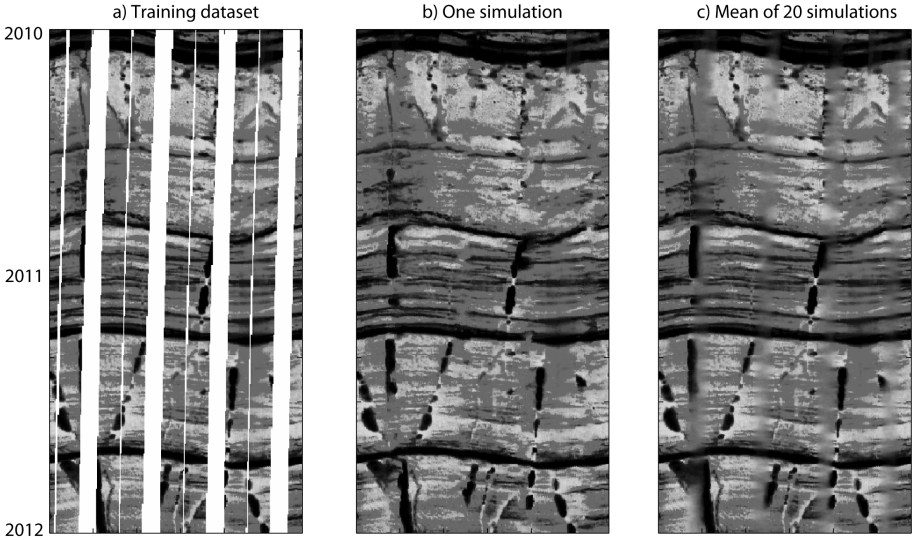



**Figure 12** a) Training dataset consisting of microresistivity measurements taken by the 8 sensors of a FMI device (courtesy of HEF Petrophysical Consulting). The vertical axis represents the borehole depth. b) One reconstruction realization. c) Mean of 20 reconstruction realizations.

## 4. Discussion and conclusion

In this paper, we presented a method for reconstructing partial images. The method can be used for a very wide range of applications. Indeed, this problem is very general even if we mainly presented 2D and 3D Earth Sciences applications. We have shown several possible applications of the method, but many others can be envisioned. For example, it is frequent that time series observations are incomplete because an instrument has been damaged. When this occurs, one usually has to build

a model to reconstruct the missing data. With the proposed method, this can be done directly and very rapidly even if the patterns and statistical distributions of the measurements are complex. Therefore, it is clear that the proposed method is versatile enough to be applied to a very wide range of problems.

The proposed algorithm propagates the properties of a known data set to locations where the values of the variable at stake are unknown. This method is based on the technique of multiple-points simulations by DS that presents several advantages over traditional methods.

First, it uses multiple-points statistics, and therefore can propagate whole patterns rather than two-points correlations. Second, because the neighborhoods are flexible, it is possible to use patterns that are combinations of irregularly positioned data, such as a set of randomly positioned points or different cross-sections. Third, it can accommodate either categorical or continuous variables. And fourth, there are no problems of data relocation as multiple-grids are not necessary.

Compared to variogram-based techniques, the presented reconstruction method has the major advantage that it needs very little parameterization. Because the spatial model is non-parametric and is given in the training data, no model inference is needed. Similarly, non-stationarity is directly accounted for if it is present in the training data and trends do not need to be known in advance. A first gain is the reduction of the uncertainty associated to the model adjustment. Another major benefit of this easy parameterization is that it minimizes the time spent adjusting the parameters for a geological model. This can be of great appeal for real-case applications where time and money are key factors.

In the different examples presented, the method successfully reconstructed partial datasets made of continuous unknown zones, such as holes in an otherwise complete image or completely informed cross-sections in a 3D domain. However, reconstruction based on very scattered datasets remains a challenge due to the lack of information at different scales. If measurement errors are present in the training dataset, the reconstructed images can be adversely affected. Moreover, the available dataset must be sufficiently large to contain a large enough diversity of patterns.

A possible direction for further research could be to use an external complete training image when the training dataset is insufficient. For each simulated grid node, the TD would first be scanned. Only in case of no matching data event the TD, the training image would then be scanned. It could allow the modeler to introduce extreme values and specific connectivity patterns not present in the TD. As for classic multiple-points simulation, the training image may be derived from expert's

interpretation and qualitative knowledge. The TI would then be a prior model, as opposed to the implicit spatial model based on the TD.

## 5. References


Arpat, B., and Caers, J. (2007), *Conditional Simulations with Patterns*, Mathematical Geology, 39, 2, 177-203.

Bersezio, R., Giudici, M., and Mele, M. (2007), *Combining sedimentological and geophysical data for high resolution 3-d mapping of fluvial architectural elements in the Quaternary Po plain (Italy).* Sedimentary Geology, 202, 230-247.

Carle, S. F., and Fogg, G. E. (1997), *Modeling spatial variability with one and multi-dimensional continuous Markov chains*, Mathematical Geology, 7, 29, 891-918.

Cornacchiulo, D., and Bagtzoglou, C. (2004), *Geostatistical Reconstruction of Gaps in Near-Surface Electrical Resistivity Data*, Vadose Zone Journal, 3, 1215-1229.

De Marsily, G., Delay, F., Gonçalvès, J., Renard, P., Teles, V., and Violette, S. (2005), *Dealing with spatial heterogenity*, Hydrogeology Journal, 13, 1, 161-183.

dell'Arciprete, D., Felletti, F., and Bersezio, R. (2008), *Simulation of fine-scale heterogeneity of meandering river aquifer analogues: comparing different approaches*, paper presented at geoENV 2008, Southampton, 8-10 September, 2008.

Deutsch, C., and Journel, A. (1992), *GSLIB: Geostatistical Software Library*, Oxford Univ. Press.

Deutsch, C., and Tran, T. (2002), *FLUVSIM: a program for object-based stochastic modeling of fluvial depositional systems*, Computers & Geosciences, 2002, 28, 525-535.

Eaton, T. (2006), *On the importance of geological heterogeneity for flow simulation*, Sedimentary Geology, 184, 187-201.

Fiorucci, P., La Barbera, P., Lanza, L., and Minciardi, R. (2001), *A geostatistical approach to multisensor rain field reconstruction and downscaling*, Hydrology and Earth System Sciences, 5, 2, 201-213.

Gómez-Hernández, J. J., and Wen, X.-H. (1998), *To be or not to be multi-gaussian? A reflection on stochastic hydrogeology*, Advances in Water Resources, 21, 1, 47-61.

Goovaerts, P. (1997), *Geostatistics for Natural Resources evaluation*, Oxford University Press, Oxford.

Guadagnini, L., Guadagnini, A., and Tartakovski, D. (2004), *Probabilistic reconstruction of geologic facies*, Journal of hydrology, 294, 57-67.

Guardiano, F., and Srivastava, M. (1993), *Multivariate geostatistics: Beyond bivariate moments*, in *Geostatistics-Troia*, pp. 133-144, Kluwier Academic.

Isaaks, E. (1984), *Indicator simulation: Application to the simulation of a high grade uranium mineralization*, in *Geostatistics for Natural Resources Characterization, Part 2*, pp. 1057-1069, D. Reidel Publishing Company.

Journel, A., and Isaaks, E. (1984), *Conditional indicator simulation: Application to a Saskatchewan deposit*, Mathematical Geology, 16, 7, 685–718.

Journel, A., and Zhang, T. (2006), *The Necessity of a Multiple-Point Prior Model*, Mathematical Geology, 38, 5, 591-610.

Kerrou, J., Renard, P., Hendricks-Franssen, H.-J., and Lunati, I. (2008), *Issues in characterizing heterogeneity and connectivity in non-multi-Gaussian media*, Advances in Water Resources, 31, 1, 147-159.

Khoshbakht, F., Memarian, H., and Mohammadnia, M. (2009), *Comparison of Asmari, Pabdeh and Gurpi formation's fractures, derived from image log*, Journal of Petroleum Science and Engineering, 67, 1-2, 65-74.

Kitanidis, P. (1986), *Parameter uncertainty in estimation of spatial functions: Bayesian analysis*, Water Resour. Res., 22, 4, 499-507.

Koltermann, C., and Gorelick, S. (1996), *Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches*, Water Resour. Res., 32, 9, 2617-2658.

Kostic, B., and Aigner, T. (2007), *Sedimentary architecture and 3-D ground penetrating radar analysis of gravelly meandering river deposits (Neckar valley, SW Germany)*, Sedimentology, 54, 789-808.

Kumar, D., and Ahmed, S. (2008), *Reconstruction of Water Level Time Series in an Aquifer Using Geostatistical Technique* in *Groundwater Dynamics in Hard Rock Aquifers*, pp. 191-200, Springer Netherlands.

Lantuéjoul, C. (2002), *Geostatistical simulation. Models and algorithms.*, Springer.

Lefranc, M., Beaudoin, B., Chiles, J.-P., Guillemot, D., Ravenne, C., and Trouiller, A. (2008), *Geostatistical characterization of Callovo–Oxfordian clay variability from high-resolution log data* Physics and Chemistry of the Earth, Parts A/B/C, 33, Supplement 1, S2-S13.

Marache, A., Riss, J., Gentier, S., and Chiles, J.-P. (2002), *Characterization and reconstruction of a rock fracture surface by geostatistics*, International Journal for Numerical and Analytical Methods in Geomechanics, 26, 9, 873-896.

Mariethoz, G., Renard, P., Cornaton, F., and Jaquet, O. (2009), *Truncated plurigaussian simulations to characterize aquifer heterogeneity*, Ground Water, 47, 1, 13-24.

Mariethoz, G., Renard, P., and Straubhaar, J. (submitted), *The direct sampling method to perform multiple-points simulations*, Water Resour. Res.

Miall, A. (1996), *The Geology of Fluvial Deposits*, Springer, New York.

Mosley, P. (1982), *Analysis of the Effect of Changing Discharge on Channel Morphology and Instream Uses in a Braided River, Ohau River, New Zealand*, Water Resour. Res., 18, 4, 800-812.

Nichols, G. (1999), *Sedimentology and Stratigraphy*

Okabe, H., and Blunt, M. (2004), *Multiple-point Statistics to Generate Pore Space Images* in *Geostatistics Banff 2004*, pp. 763-768, Springer Netherlands.

Ortiz, J. M., and Deutsch, C. V. (2004), *Indicator Simulation Accounting for Multiple-Point Statistics*, Mathematical Geology, 36, 5, 545-565.

Renard, P., Gómez-Hernández, J., and Ezzedine, S. (2005), *Characterization of Porous and Fractured Media*, in *Encyclopedia of Hydrological Sciences*, John Wiley & Sons.

Rubin, Y., Lunt, I., and Bridge, J. (2006), *Spatial variability in river sediments and its link with river channel geometry*, Water Resour. Res., 42, W06D16.

Saripally, I. (2008), *Evaluating Data Conditioning Accuracy of MPS Algorithms and the Impact on Flow Modeling*, paper presented at 21th SCRF Meeting, Stanford University, May 8-9, 2008.

Straubhaar, J., Walgenwitz, A., Renard, P., and Froidevaux, R. (2008), *Optimization issues in 3D multipoint statistics simulation*, paper presented at Geostats 2008, 1-5 Dec. 2008, Santiago, Chile.

Strebelle, S. (2002), *Conditional simulation of complex geological structures using multiple point statistics.*, Mathematical Geology, 34, 1, 1-22.

Strebelle, S., Payrazyan, K., and J., C. (2002), *Modeling of a Deepwater Turbidite Reservoir Conditional to Seismic Data Using Multiple-Point Geostatistics*, paper presented at the 2002 SPE Annual Technical Conference and Exhibition, San Antonio, September 29 - October 2, SPE paper 77425.

Wu, H., and Pollard, D. (2002), *Imaging 3-D fracture networks around boreholes*, AAPG Bulletin, 86, 4, 593-604.

Wu, J., Boucher, A., and Zhang, T. (2008), *A SGeMS code for pattern simulation of continuous and categorical variables: FILTERSIM*, Computers & Geosciences, 34, 12, 1863-1876.

Youngseuk, K., Mukerji, T., and Amos, N. (2004), *Permeability prediction from thin sections: 3D reconstruction and Lattice-Boltzmann flow simulation*, Geophys. Res. Lett., 31, L04606.

Zappa, G., Bersezio, R., Felletti, F., and Giudici, M. (2005), *Modeling heterogeneity of gravel-sand, braided stream, alluvial aquifers at the facies scale* Journal of hydrology, 325, 1-4, 134-153.

Zhang, T., Switzer, P., and Journel, A. (2006), *Filter-Based Classification of Training Image Patterns for Spatial Simulation*, Mathematical Geology, 38, 1, 63-80.

# Chapter 4

# Super-resolution using multiple-points statistics*

*"Don't believe what your eyes are telling you. All they show is limitation. Look with your understanding."*

*Richard Bach, Jonathan Livingston Seagull*

* This chapter has been submitted for publication in IEEE Transactions for Image Processing as:
Mariethoz, G., P. Renard. Super-Resolution Using Multiple-Points Statistics.
The algorithms described in this article are protected by the international patent 2008WO-EP009819.

**Abstract** The resolution of measurement devices can be insufficient for certain purposes. Super-resolution techniques aim at simulating parameters values at a sub-pixel scale. We propose to use the geostatistical method of multiple-points by Direct Sampling (DS) to stochastically simulate the structures at scales smaller than the measurement resolution. These structures are inferred using a hypothesis of scale-invariance on the spatial patterns found at the coarse scale. The proposed multiple-point super-resolution method is able to deal with various kinds of variables (e.g. continuous or categorical) and can be extended to multivariate problems. The approach is illustrated with examples of satellite imaging and digital photography.

## 1. Introduction

With modern computers able to deal with high definition models, it is possible to describe and predict large and small scale physical phenomena. Such fine models are much needed in various fields of Earth Science, such as for example Hydrology, Hydrogeology and Meteorology (Von Storch, et al., 1993; R.Wilby and Wigley, 1997; Ferraris, et al., 2003; Liu, et al., 2008).

Most models are fed by data coming from various kinds of measurement devices, whose sampling resolution is limited (Bertero and Boccacci, 2003). Hence, the measurement scale is often larger than the scale at which physical processes occur, making the understanding of small scale phenomena a challenging issue, even when important datasets are available (Schulze-Makuch and Cherkauer, 1998; Zlotnik*, et al.*, 2000). This situation is not likely to change because no matter the spatial resolution of the data, some limitation on the measurement scale will always remain. Therefore, finding methods to deal with coarse sampling is mandatory. The problem treated in the present paper is the case where the variable of interest is exhaustively known, but the spatial resolution of this information is too coarse. There is a missing scale where no information is available. This problem is common to the disciplines of geostatistics and image processing. One could think that information on a scale smaller than the measurement scale cannot be inferred as it is simply not available in the data. However, methods have been developed to address this problem, which is known as Super-Resolution in the Image Processing literature (Baker and Kanade, 2002; Park, et al., 2003).

Simulating the structures smaller than the coarse pixel size requires inventing them in some way, or borrowing them from proxys or from a training image (TI) which is assumed representative of the uninformed scale.

For hydrogeological models, efficient Super-Resolution can be performed through an inverse procedure (Wen, et al., 1997; Grimstad, et al., 2003) relying on indirect data such as for example production history from wells or pumping tests data when the variable of interest is hydraulic conductivity. Such problems, often known as downscaling, include other constraints than the small-scale spatial structure only. For example when downscaling hydraulic conductivity, the local mean conductivity of a downscaled area must be equivalent to the respective coarse-scale values. In the frame of 2-points geostatistics, techniques such as block simulation (Liu and Journel, 2008) are available to account for these constraints. The above mentioned methods are efficient for specific hydrogeological problems where data are available as a base for the inverse procedure. In this paper, we do not present a method for downscaling, but for super-resolution. Therefore, we do not impose such constraints.

Land cover estimation based on coarse remote sensing images is a common super-resolution problem (Nguyen*, et al.*, 2006). Tatem*, et al.* (2002) estimate the class composition of the coarse pixels as well as the spatial distribution of these classes within the pixels. Super-resolution is formulated as an optimization problem that is solved using a Hopfield neural network. Constraints are imposed on the spatial structures at sub-pixel scale, given by variograms derived from proxys.

In Image Processing, number of methods and algorithms are devoted to Super-Resolution, for both continuous and categorical variables (See a review in Farsiu*, et al.*, 2004). Excellent results in this domain are given by TI-based techniques (Atkins*, et al.*, 1999; Freeman*, et al.*, 2002) that derive the relationship between small-scales features and their coarse counterparts from a database of high-resolution TIs and their coarsened/blurred versions. Other methods (Tsai and Huang, 1984; Sroubek, et al., 2007) fuse a sequence of low-resolution images (e.g. successive frames of a movie scene) to produce a higher-resolution image.

Geostatistical methods are a very advanced and appropriate way of dealing with spatial uncertainty (De Marsily*, et al.*, 2005; Journel and Zhang, 2006). Among them, Multiple-Points (MP) simulation is a recent family of stochastic image generation methods used in Earth Sciences applications and especially in petroleum engineering (Caers, 2005; Hu and Chugunova, 2008). It is a valuable tool for integrating an empirical and non-parametric structural concept in stochastic models. This non-parametric model takes the shape of a training image (TI) that depicts the spatial features of the variable of interest. The principle of MP simulations is to recompose the patterns found in the TI to obtain a field respecting various constraints, such as conditioning data. Numerous MP algorithms have been

developed (Guardiano and Srivastava, 1993; Strebelle, 2002; Zhang, et al., 2006; Arpat and Caers, 2007; Straubhaar, et al., 2008; Mariethoz, et al., submitted). A stochastic formulation of Super-Resolution using geostatistical methods was investigated by Boucher et al. (2008) using multiple-points geostatistical methods. In this paper, we show that these multiple-points methods, and especially the Direct Sampling (DS) (Mariethoz, et al., submitted), can be a precious aid for addressing the issue of increasing the resolution of existing images.

Despite the numerous existing techniques, there is still a need for generic Super-Resolution algorithms capable of producing satisfying high-resolution images using one coarse image only (as opposed to methods using additional information to inform the fine scale, such as TIs, series of coarse images or statistics derived from a proxy). In this context, we propose to perform Super-Resolution using multifractal cascades (Mandelbrot, 1974) combined with MP simulations for reconstructing the fine scale structures. The specificity of our approach is that the algorithm does not need a TI or a proxy, because we infer directly MP statistics from the coarse image and we use them at the finer, unknown scale. This way, the small scale structures are borrowed from the large-scale ones, using a hypothesis of self-similarity. This algorithm is straightforward in terms of implementation and parameterization, and is robust as it makes very few assumptions on the missing scale. Moreover, implemented in a stochastic framework, it allows for a complete probabilistic solution of Super-Resolution problems.

The first part of the paper presents the DS algorithm, which is the main tool for solving the considered super-resolution problems. The second section describes the super-resolution algorithm and illustrates it with examples.

## 2. The direct sampling algorithm

Many geostatistical algorithms are devoted to produce realizations of a spatially correlated variable $Z$ at all $N$ locations $\mathbf{x}_i$ of a regular grid, with $i = [1, \cdots, N]$. Each such realization is a sample of the $N$-dimensional joint distribution

$$\mathrm{F}(z, \mathbf{x}) = \mathrm{Prob}\left\{Z(\mathbf{x}_1) \le z,\ Z(\mathbf{x}_2) \le z, \cdots, Z(\mathbf{x}_3) \le z\right\}. \quad (1)$$

Sequential simulation algorithms (Deutsch and Journel, 1992) are a practical way of sampling (1) by performing the following decomposition:

$$\mathrm{F}(z,\mathbf{x}) = \mathrm{Prob}\left\{Z(\mathbf{x}_1) \le z\right\} \cdot \mathrm{Prob}\left\{Z(\mathbf{x}_2) \le z \mid z(\mathbf{x}_1)\right\} \cdot \ldots \cdot \mathrm{Prob}\left\{Z(\mathbf{x}_m) \le z \mid z(\mathbf{x}_1), \ldots, z(\mathbf{x}_{m-1})\right\}.$$

(2)

Sequential simulations proceed by considering only a limited neighborhood of size $n$, with $n<<N$ to limit computational burden. At each location $\mathbf{x}$, the lag vectors $\mathbf{L}=\{\mathbf{h}_1,\ldots,\mathbf{h}_n\}=\{\mathbf{x}_1-\mathbf{x},..,\mathbf{x}_n-\mathbf{x}\}$ define the data event at location $\mathbf{x}$, $\mathbf{d}_n(\mathbf{x},\mathbf{L})=\{Z(\mathbf{x}+\mathbf{h}_1),...,Z(\mathbf{x}+\mathbf{h}_n)\}$. Given this data event, sequential simulation techniques find the ccdf for the variable of interest $Z$:

$$\mathrm{F}(z,\mathbf{x},\mathbf{d}_n)=\mathrm{Prob}\{Z(\mathbf{x})\le z\,|\mathbf{d}_n(\mathbf{x},\mathbf{L})\}. \tag{3}$$

Once the ccdf (3) is determined, a value for $Z(\mathbf{x})$ is drawn from it, and is thereafter considered as conditioning data when simulating the remaining nodes.

The DS algorithm (Mariethoz*, et al.*, in press) is a geostatistical sequential simulation algorithm making use of MP statistics. MP simulations algorithms such as SNESIM (Strebelle, 2002), FILTERSIM (Zhang*, et al.*, 2006) or IMPALA (Straubhaar*, et al.*, 2008) derive the conditional distributions (4) from a TI given as a model and representing the desired spatial structure of the variable of interest (Strebelle, 2002; Journel and Zhang, 2006). The particularity of DS lies in the fact that it does not store probabilities associated to all pixels configurations found in the TI. Instead, at each simulated node $\mathbf{x}$, the TI is sampled until the appropriate pattern is found. For each of the successive samples, the distance or mismatch $d\{\mathbf{d}_n(\mathbf{x},\mathbf{L}),\mathbf{d}_n(\mathbf{y},\mathbf{L})\}$ between the data event observed in the simulation $\mathbf{d}_n(\mathbf{x,L})$ and the one sampled from the training image $\mathbf{d}_n(\mathbf{y,L})$ is calculated ($\mathbf{y}$ denotes the nodes of the TI). If there is no mismatch or if it is lower than a given threshold, the sampling process is stopped and the value at the central node of the data event in the TI $Z(\mathbf{y})$ is directly pasted in the simulation at the location $\mathbf{x}$. At no point is the ccdf (3) is needed.

Using different measures of distance offers a high degree of flexibility. It is therefore possible to accommodate categorical as well as continuous variables. For categorical variables, the distance between a data event found in the simulation and another one found in the TI $d\{\mathbf{d}_n(\mathbf{x},\mathbf{L}),\mathbf{d}_n(\mathbf{y},\mathbf{L})\}$ is given by the proportion of non-matching nodes. It is calculated by using the indicator variable $a$ that equals 0 if two nodes have identical value and 1 otherwise:

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}=\frac{1}{n}\sum_{i=1}^{n}a_i \quad \in[0,1], \quad \textit{where } a_i=\begin{cases}0 & \textit{if } Z(\mathbf{x}_i)=Z(\mathbf{y}_i)\\ 1 & \textit{if } Z(\mathbf{x}_i)\neq Z(\mathbf{y}_i)\end{cases}. \tag{4}$$

A convenient measure of distance able to accommodate continuous variables is the normalized pair wise Manhattan distance:

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|Z(\mathbf{x}_i)-Z(\mathbf{y}_i)\right|}{\text{argmax}(Z^{ti})-\text{argmin}(Z^{ti})}\in[0,1]. \quad (5)$$

Simulations resulting from DS reproduce the spatial properties of the scanned image. Moreover, provided that the TI is large enough, these properties are reproduced even if they show a high degree of complexity because the method exploits high order statistics from the TI. Contrarily to other MP methods, DS does not need multiple-grids to reproduce structures at different scales. Instead, it accommodates data events of different sizes. As such, it does not encounter problems of relocation of conditioning data on the different simulation grids (Saripally, 2008), making it easy to accommodate large data sets. As reconstruction problems rely on large datasets, this feature is a major advantage.

Another feature of the DS is the definition of multivariate data events, allowing to process multivariate images. At each grid node, $Z$ is a vector of size $m$, where $m$ is the number of variables. The conditional cumulative density function (3) for the variable $Z_k$ is then expressed as

$$\mathrm{F_k}(z,\mathbf{x},\mathbf{d}^1_{n_1},\cdots,\mathbf{d}^m_{n_m})=\text{Prob}\left\{Z_k(\mathbf{x})\le z\,|\,\mathbf{d}^1_{n_1},\cdots,\mathbf{d}^m_{n_m}\right\},\quad k=1,...,m. \quad (6)$$

Each variable $Z_k$ involved in the multivariate analysis can have a different neighborhood and a specific data event $\mathbf{d}^k_{n_k}(\mathbf{x},\mathbf{L}^k)=\left\{Z_k(\mathbf{x}+\mathbf{h}^k_1),...,Z_k(\mathbf{x}+\mathbf{h}^k_{n_k})\right\}$. The distance between a joint data event found in the simulation and one found in the TI is defined as a weighted average of the individual distances defined previously.

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\}=\sum_{k=1}^{m}w_k\,d\left\{\mathbf{d}^k_{n_k}(\mathbf{x},\mathbf{L}^k),\mathbf{d}^k_{n_k}(\mathbf{y},\mathbf{L}^k)\right\}\in[0,1],\ \text{with}\ \sum_{k=1}^{m}w_k=1\ \text{and}\ w_k\ge 0. \quad (7)$$

The weights $w_k$ are defined by the user. This allows accounting for the fact that the pertinent measure of distance may be different for each variable. Joint simulations are performed using a single (random) path that visits all components of vector $Z$ at all nodes of the SG.

By accounting for multiple-points correlations between variables, DS allows respecting cross-correlations between all combinations of nodes within multivariate data events. An example of multivariate super-resolution is shown below.

## 3. Dealing with a missing scale

We propose a general method that does not rely on any other input data than the coarse image and that has the advantage of being very parsimonious in parameterization. It assumes that the parameters of interest have properties of scale invariance (Mandelbrot, 1967), thus allowing to infer the small scale structures from the large ones. This assumption constitutes a model for the missing scale. The outcomes of physical processes can often present fractal properties. For example, self-similar characteristics have been observed for a wide range of geological physical properties (Kiraly, 1988; Barton and La Pointe, 1991; Turcotte, 1992). Taking profit of these fractal properties, it becomes possible to perform Super-Resolution of coarse images.

We propose to accomplish Super-Resolution in a similar way as the fractal cascades proposed by Mandelbrot (1974). The main idea is to consider a coarse image as a sampling of the same image in high-resolution. At each step of a 2D Super-Resolution problem, the coarse nodes are divided into four children nodes. The value of each coarse node is assigned to one of its children nodes. The exact location of the value on the fine grid is unknown. Hence, the child node that receives its parent value is chosen randomly. Once this migration has been accomplished, the values of the three remaining children nodes are determined using DS. Here, instead of determining the value of the children nodes by scanning a given TI, the coarse image is scanned, even if it has a different resolution than the simulation grid (all pixels are 4 times larger on the coarse image). The original coarse grid is used as training image. Treating a 3D problem is identical, except that each coarse grid node would have 8 children nodes.

Traditional MP techniques could also be used instead of DS. Using the usual multiple-grids approach (Strebelle, 2002), it would be accomplished by creating an additional fine multiple-grid and assuming that the patterns catalogue (the search tree or the data events list) remains unchanged between this finer grid and the coarser ones. Nevertheless, problems would arise when migrating large numbers of conditioning data between multiple-grids, because several conditioning data may correspond to one coarse multiple-grid node, resulting in approximations on the simulation of the large structures. Moreover, super-resolution would then be limited to categorical variables.

The Super-Resolution algorithm proceeds as follows:

1. Given a coarse image on grid $G_1$ of size $[X,Y]$, create a finer grid $G_2$ of size $[2X,2Y]$.
2. For each node $\mathbf{x}^{G1}$ of $G_1$, assign $Z(\mathbf{x}^{G1})$ to one of its 4 children nodes on $G_2$ (chosen randomly).
3. Define a random path through all nodes of $G_2$ whose value has not already been assigned.
4. Simulate the remaining children nodes $Z(\mathbf{x}^{G2})$ with some MP technique (DS in the present case), using $G_1$ as training image grid.

This Super-Resolution procedure can be performed many times, using $G_2$ as coarse grid and $G4$ (size $[4X,4Y]$) as fine grid, etc. Figure 1 illustrates two iterations of Super-Resolution for a single coarse grid node.

Note that contrary to downscaling techniques, this algorithm does not ensure that the mean of children nodes values is equal to the value of their respective parent nodes.

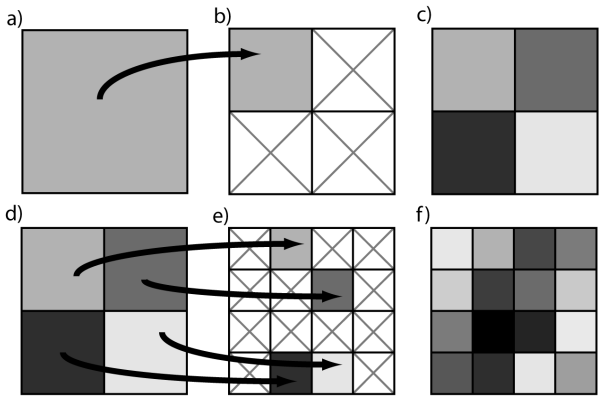



**Figure 1** Visual representation of the MP Super-Resolution procedure for a single coarse grid node. a) and b) Migrate the value of the coarse grid to one of the 4 possible equivalent fine grid nodes. c) Simulate the remaining 3 fine grid nodes using the coarse grid as training image. d), e) and f) Second Super-Resolution iteration.

The first example illustrates super-resolution applied to an image that was built by applying geostatistical methods on an aerial photograph of braided channels in the Ohau River, New Zealand (Mosley, 1982). The result is a grid populated with spatially correlated, anisotropic and non-Gaussian, transmissivity (*T*) values. Kerrou, *et al.*(2008) explains in details the procedure used for generating the distribution of transmissivity values. The image size is 1000 m by 400 m and it has been discretized in 440x176 grid nodes.

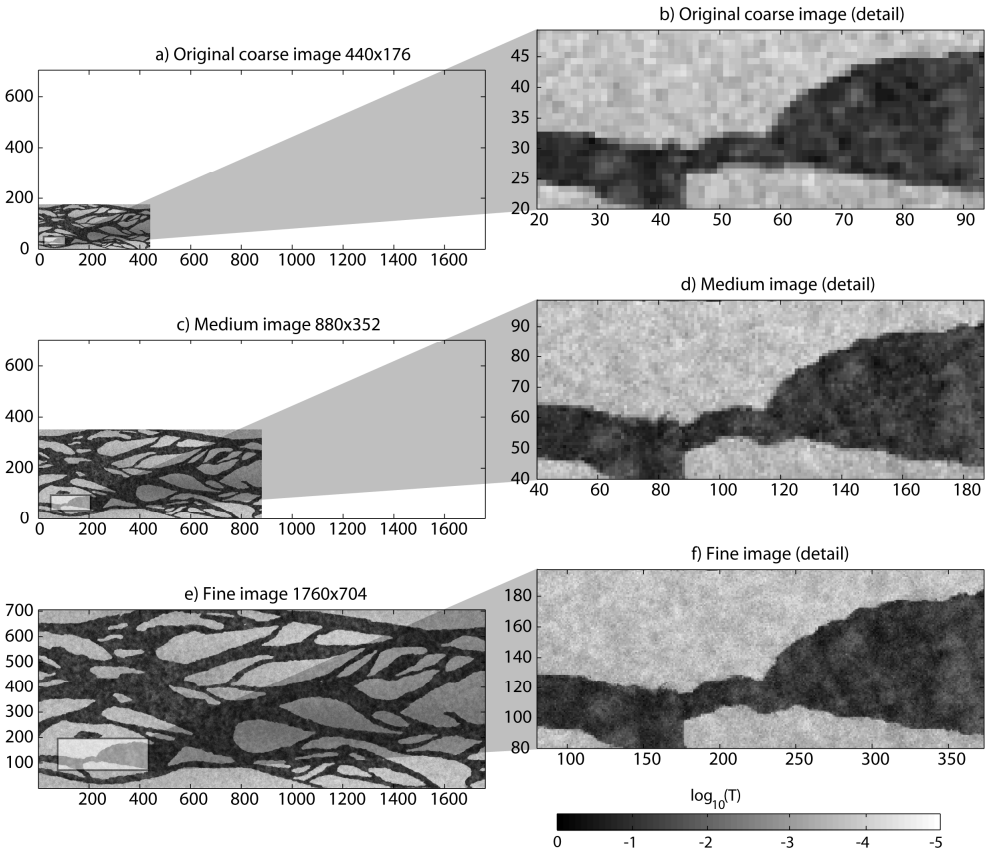



**Figure 2** Results of two Super-Resolution steps for a continuous variable. a) Original coarse image (G1), which is the only available input data for the Super-Resolution algorithm. b) A detail of the coarse image. c) Resulting image after one Super-Resolution iteration (G2). d) A detail of the medium image. e) Resulting image after two Super-Resolution iterations (G3). f) A

detail of the fine image. Note that the axes do not represent the real system of coordinates, but the number of cells in each image.

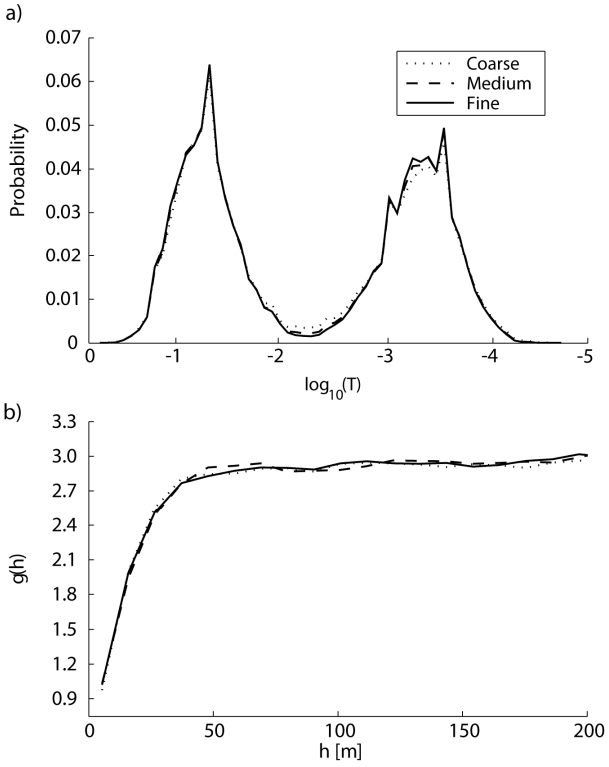



**Figure 3** a) Histograms of the values of coarse, medium and fine images. b) Omnidirectional variograms of the three images. Variograms have been computed using the real system of coordinates, with a domain of size 1000x400 meters in the three cases.

Two Super-Resolution steps are accomplished with the DS algorithm, using the measure of distance (5) because the variable is continuous. Figure 2a shows the original coarse image $G_1$. Figure 2b depicts a detail of $G_1$, with coarse pixels clearly visible. The Super-Resolution procedure can be performed many times, using each simulation from the previous step as a new coarse image that is sampled again. Figure 2c to 2f display the results of two Super-Resolution iterations. In the detailed zone, the image becomes sharper at each step while preserving the structures present in the coarse image. Comparison of the histograms of the three images (Figure 3a)

show that Super-Resolution does not induce a bias in the distribution of transmissivity values. The variograms (Figure 3b) clearly indicate that the spatial structure of the downscaled field is not affected. Histograms and Variograms are perfectly reproduced, without performing any histogram transform or variogram adjustment. Fractal dimensions do not need to be calculated, nor is any adjustment of parameters necessary, except for the parameters of the DS algorithm. The model for inferring the fine-scale structure is the hypothesis of self-similarity applied to the patterns of the coarse image.

We discussed above how DS allows performing multivariate simulations with variables that can present any kind of multiple-points correlations. We use this feature to consider the three basic color channels as joint variables and to perform super-resolution on them simultaneously. The original RGB image (Figure 4a) is first decomposed in three variables, one for each color channel (Figure 4b). These variables are randomly migrated on a finer grid (Figure 4c), as described above, to constitute the conditioning data for a multivariate simulation of the three color variables using DS. The coarse image (Figure 4a) is used as TI. 100 super-resolution simulations are generated (3 of them are shown in Figure 4d) each having different randomly migrated conditioning data.

The resulting stack of super-resolution realizations provides a complete probability distribution for each fine pixel. This probabilistic result can be exploited in different ways depending on the field of application. For example, one can choose a single preferred realization if the textures have to be preserved. The mean of all realizations (calculated on each color channel) can give a smoothed estimation (Figure 4e), and the mode (Figure 4f) gives the most probable color at each pixel. Uncertainty on the resulting images can be evaluated by mapping the standard deviation, averaged on the three colors channels (Figure 4g). The locations with high uncertainty are the ones populated with patterns having few replicates in the coarse image. These are the pixels patterns with sharp contrasts such as areas left and right of the nose, and more generally the lineaments of the face.

A second super-resolution step is shown in Figure 5. Each simulation from the previous step is sampled again (Figure 5a), resulting in finer realizations (Figure 5b) on which statistics can be computed (Figure 5c to 5e). The images resulting from the second step show sharp contrasts between color tones. This can be explained by the lack of color diversity in the original coarse image. As the DS algorithm only copies color values from the TI, it cannot create new tones (as opposed to interpolation methods that create intermediate colors between the existing ones). Using a larger

and richer original coarse image would probably mitigate the appearance of such artifacts by providing a larger pool of color tones and patterns to use.

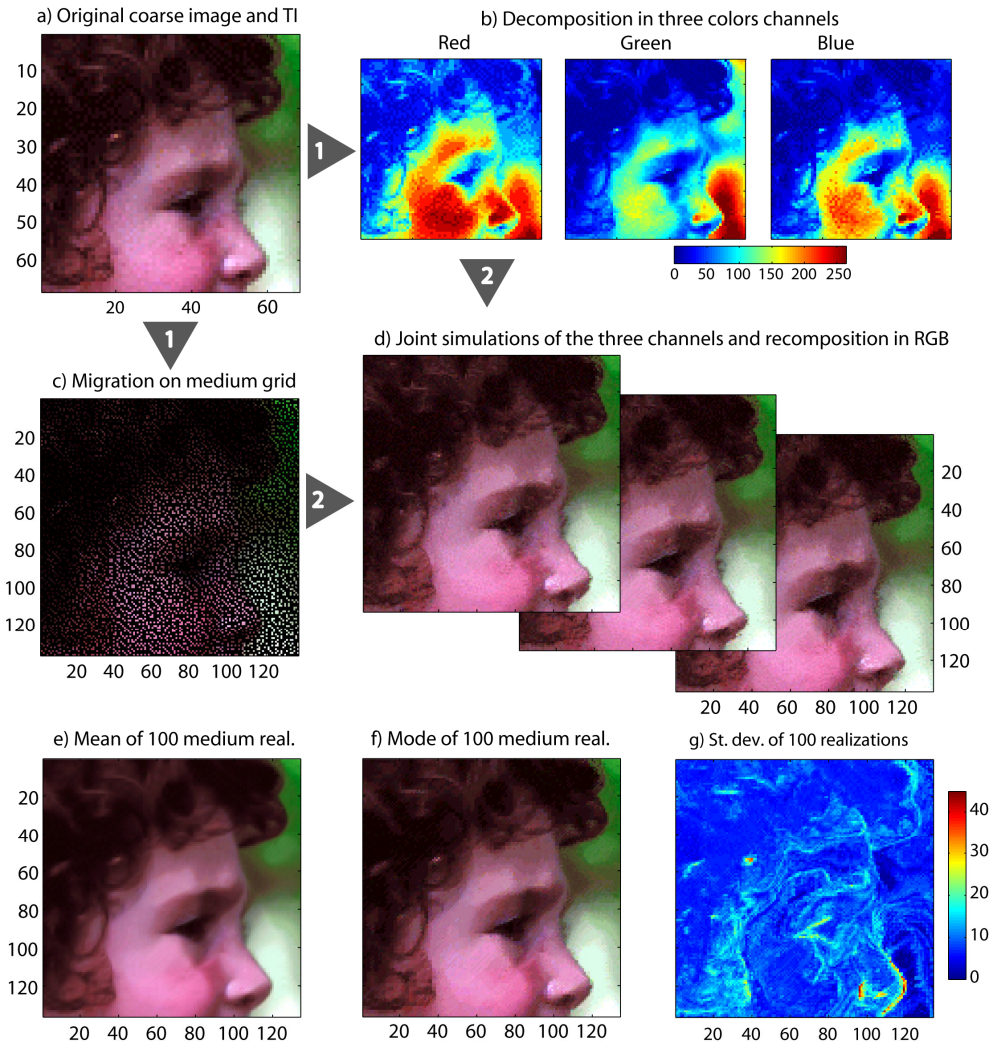



**Figure 4** Illustration of Super-Resolution using DS with a small, colorized non-stationary image (Freeman, et al., 2000). a) The coarse image is the only input to the Super-Resolution algorithm (70 by 70 pixels). It is made of the three color variables shown in b). c) Migration to finer grid. d) Images resulting from recomposition of the three jointly simulated colors channels (140 by 140 pixels). Different use of the realizations stack can include

computing the mean e), the mode f) or the standard deviation of the entire stack (100 realizations).

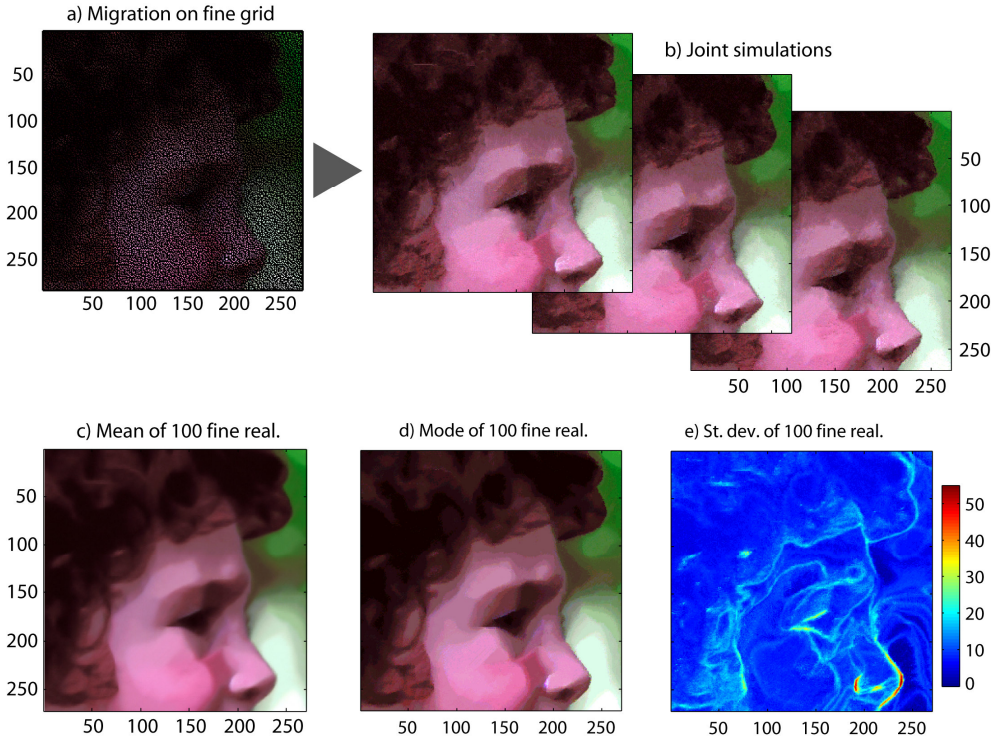



**Figure 5** The Super-Resolution procedure is repeated by a) sampling each lower resolution simulation and using it as TI to obtain finer realizations b) (280 by 280 pixels). c) to e) Statistics on the stack of 100 fine realizations.

The Super-Resolution simulations are satisfying in the sense that they create realistic features at the small, unknown scale. Whereas usual interpolation techniques create artifacts at small scale such as blurry edges (see for example cubic spline interpolation on Figure 6), DS Super-Resolution preserves the sharpness of lineaments and the different textures such as skin and hair. For instance, note the hair-like texture of the eyebrows in Figure 5b. No such texture is visible for the eyebrows on the original coarse image (Figure 4a). The algorithm borrowed patterns coming from the hair and used them for the eyebrows of the finer grids, thus rendering a realistic texture.

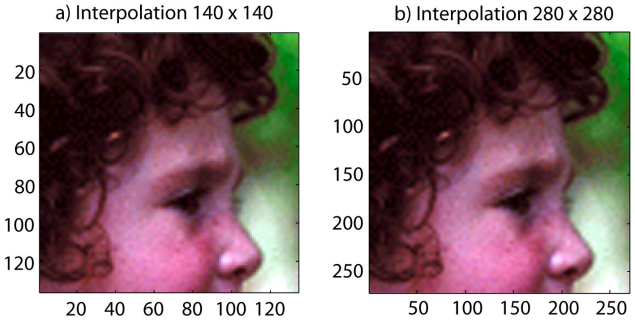



**Figure 6** Super-Resolution on the medium (a) and fine (b) grids using cubic spline interpolation (Adobe Photoshop 9).

In geostatistics, the non-stationarity of images is usually addressed with specific methods. Here, the non-stationarity of the image is correctly handled without using such extra processing. This is possible because the migration on the finer grids produces a dense enough sampling that accurately describes the location of the major textural zones (hair, skin, green or white background).

## 4. Conclusion

In this paper, we present a method for performing Super-Resolution of coarse images.

Super-Resolution of coarse images poses the problem of inferring structures at a scale where no information is available. Earth Sciences propose methods for addressing this problem relying on indirect data. Image Processing methods rely on databases of TIs or on many instances of the same low-resolution scene. Our method infers information on the missing scale by making the assumption of fractal properties in the patterns of the variable to simulate. Provided that the assumption of patterns self-similarity holds, the resulting fine images present the same properties as their coarse counterparts, both in terms of visual aspect and reproduction of spatial statistics. This Super-Resolution method is accurate, parsimonious in parameterization and valid over a wide range of scales. The method does not need explicit prior models such as training image(s) or variogram(s). Nevertheless, if the coarse image is very small, it may not contain a very diverse spectrum of patterns, thus resulting in artifacts when inferring the small scale features.

The possibility to perform multivariate super-resolution has been illustrated with the 3 RGB channels of a color image.

An aspect of the method that can be further developed is that the random migration of the coarse pixels on the fine grid can result in patterns that are incompatible with the patterns found in the coarse grid. This aspect of the method could be addressed by designing a smart migration strategy that would maximize the compatibility of the patterns after migration with the patterns of the coarse image.

Compared to super-resolution techniques used in image processing, the proposed Super-Resolution method is limited to self-similar images, but this hypothesis is loose as shown by the last example, where the algorithm performs well even when the image is by no means entirely fractal. Because it is able to produce multiple realizations, the method has the major advantage of providing statistical information on the uncertainty of the simulated properties, which can become of utmost importance when decision variables are at stake (e.g. mapping of pollution, estimation of natural resources, etc).

The range of potential applications of this method is wide. It can be for example applied to remote sensing images (e.g. satellite imaging, geophysics) whose resolution is insufficient for given applications or time series measurements (such as precipitation intensity at a given location) that were recorded with widely spaced time steps.

A possible direction for further research is to develop the use of this super-resolution method on unstructured grids. Several challenges are involved, such as the migration from a coarse to a fine mesh and the characterization of a distance that is valid when the lag vectors of both data events (in the TI and in the SG) are substantially different.

## 5. References


Arpat, B., and Caers, J. (2007), *Conditional Simulations with Patterns*, Mathematical Geology, 39, 2, 177-203.

Atkins, C., Bouman, C., and Allebach, J. (1999), *Tree-based resolution synthesis*, paper presented at Image Processing, Image Quality, Image Capture Systems Conference (PICS-99), IS&T, Savannah, Georgia.

Baker, S., and Kanade, T. (2002), *Limits on super-resolution and how to break them*, IEEE Transactions on pattern analysis and machine intelligence, 24, 9, 1167-1183.

Barton, C., and La Pointe, P. (1991), *Fractals in the Earth Sciences*, Plenum Press.

Bertero, M., and Boccacci, P. (2003), *Super-resolution in computational imaging*, Micron, 34, 6, 265-273.

Boucher, A., Kyriakidis, C., and Cronkite-Ratcliff, C. (2008), *Geostatistical Solutions for Super-Resolution Land Cover Mapping*, IEEE Transactions on Geoscience and Remote Sensing, 46, 1, 272-283.

Caers, J. (2005), *Petroleum Geostatistics*, Society of Petroleum Engineers.

De Marsily, G., Delay, F., Gonçalvès, J., Renard, P., Teles, V., and Violette, S. (2005), *Dealing with spatial heterogenity*, Hydrogeology Journal, 13, 1, 161-183.

Deutsch, C., and Journel, A. (1992), *GSLIB: Geostatistical Software Library*, Oxford Univ. Press.

Farsiu, S., Robinson, D., Elad, M., and Milanfar, P. (2004), *Advances and Challenges in Super-Resolution*, International Journal of Imaging Systems and Technology, 14, 2, 47-57.

Ferraris, L., Gabellani, S., and Rebora, N. (2003), A comparison of stochastic models for spatial rainfall downscaling, *Water Resour. Res.*, doi: 1368.

Freeman, W., Pasztor, E., and Carmichael, O. (2000), *Learning Low-Level Vision*, International Journal of Computer Vision, 40, 1, 25-47.

Freeman, W. T., Jones, T. R., and Pasztor, E. C. (2002), *Example-Based Super-Resolution*, IEEE Computer Graphics and Applications, 22, 2, 56-65.

Grimstad, A., Mannseth, T., Nævdal, G., and Urkedal, H. (2003), *Adaptive multiscale permeability estimation*, Computational Geosciences, 2003, 7, 1-25.

Guardiano, F., and Srivastava, M. (1993), *Multivariate geostatistics: Beyond bivariate moments*, in *Geostatistics-Troia*, pp. 133-144, Kluwier Academic.

Hu, L., and Chugunova, T. (2008), *Multiple-Point Geostatistics for Modeling Subsurface Heterogeneity: a Comprehensive Review*, Water Resour. Res., 44, W11413.

Journel, A., and Zhang, T. (2006), *The Necessity of a Multiple-Point Prior Model*, Mathematical Geology, 38, 5, 591-610.

Kerrou, J., Renard, P., Hendricks-Franssen, H.-J., and Lunati, I. (2008), *Issues in characterizing heterogeneity and connectivity in non-multi-Gaussian media*, Advances in Water Resources, 31, 1, 147-159.

Kiraly, L. (1988), *Large scale 3-D groundwater flow modelling in highly heterogeneous geologic medium*, in *Groundwater Flow and Quality Modelling*, pp. 761- 775

Liu, X., Coulibaly, P., and Evora, N. (2008), *Comparison of data-driven methods for downscaling ensemble weather forecasts*, Hydrology and Earth System, 12, 2, 615-624.

Liu, Y., and Journel, A. (2008), *A package for geostatistical integration of coarse and fine scale data*, Computers & Geosciences, 35, 3, 527-547.

Mandelbrot, B. (1967), *How long is the coast of Britain? Statistical self-similarity and fractional dimension*, Science, 156, 3775, 636-638.

Mandelbrot, B. (1974), *Intermittent turbulence in self-similar cascades: Divergence of high moments and dimensions of the carrier*, J. Fluid Mech., 62, 2, 331-358.

Mariethoz, G., Renard, P., and Straubhaar, J. (in press), *The direct sampling method to perform multiple-points simulations*, Water Resour. Res.

Mariethoz, G., Renard, P., and Straubhaar, J. (submitted), *The direct sampling method to perform multiple-points simulations*, Water Resour. Res.

Mosley, P. (1982), *Analysis of the Effect of Changing Discharge on Channel Morphology and Instream Uses in a Braided River, Ohau River, New Zealand*, Water Resour. Res., 18, 4, 800-812.

Nguyen, M., Atkinson, P., and Lewis, H. (2006), *Superresolution Mapping Using a Hopfield Neural Network With Fused Images*, IEEE Transactions on Geoscience and Remote Sensing, 44, 3, 736-749.

Park, S., Park, M., and Kang, M. (2003), *Super-resolution image reconstruction: a technical overview*, Signal Processing Magazine, IEEE, 20, 3, 21-36.

R.Wilby, and Wigley, T. (1997), *Downscaling general circulation model output: a review of methods and limitations*, Progress in Physical Geography, 21, 4, 530-548.

Saripally, I. (2008), *Evaluating Data Conditioning Accuracy of MPS Algorithms and the Impact on Flow Modeling*, paper presented at 21th SCRF Meeting, Stanford University, May 8-9, 2008.

Schulze-Makuch, D., and Cherkauer, D. (1998), *Variations in hydraulic conductivity with scale of measurement during aquifer tests in heterogeneous, porous carbonate rocks*, Hydrogeology Journal, 6, 204-215.

Sroubek, F., G., C., and J., F. (2007), *A Unified Approach to Superresolution and Multichannel Blind Deconvolution*, IEEE Transactions on Image Processing, 16, 9.

Straubhaar, J., Walgenwitz, A., Renard, P., and Froidevaux, R. (2008), *Optimization issues in 3D multipoint statistics simulation*, paper presented at Geostats 2008, 1-5 Dec. 2008, Santiago, Chile.

Strebelle, S. (2002), *Conditional simulation of complex geological structures using multiple point statistics.*, Mathematical Geology, 34, 1, 1-22.

Tatem, A., Lewis, H., Atkinson, P., and Nixon, M. (2002), *Super-resolution land cover pattern prediction using a Hopfield neural network*, Remote Sensing of Environment, 79, 1-14.

Tsai, R. Y., and Huang, T. S. (1984), *Multiframe image restoration and registration*, in *Advances in Computer Vision and Image Processing*, pp. 317-339, JAI Press.

Turcotte, D. (1992), *Fractals and Chaos in Geology and Geophysics*, Cambridge Press.

Von Storch, H., Zorita, E., and Cubasch, U. (1993), *Downscaling of Global Climate Change estimates to Regional Scales: An Application to Iberian Rainfall in Wintertime*, Journal of climate, 6, 6, 1161-1171.

Wen, X., Deutsch, C., and Cullick, A. (1997), *High resolution reservoir Models Integrating Multiple-Well Production Data*, SPE, SPE 38728.

Zhang, T., Switzer, P., and Journel, A. (2006), *Filter-Based Classification of Training Image Patterns for Spatial Simulation*, Mathematical Geology, 38, 1, 63-80.

Zlotnik, V. A., Zurbuchen, B. R., Ptak, T., and Teutsch, G. (2000), *Support volume and scale effect in hydraulic conductivity: experimental aspects*, in *Theory, modelling, and field investigations in hydrogeology: A special volume in homor of Shlomo P. Neuman's 60th birthday*, pp. 215-231, Geological Society of America.

# Chapter 5

# A general parallelization strategy for random path based geostatistical simulation methods*

*"The world is changing very fast. Big will not beat small anymore. It will be the fast beating the slow."*

*Rupert Murdoch*

* This chapter has been submitted for publication in Computers & Geosciences as:
Mariethoz, G. A general parallelization strategy for random path based geostatistical simulation methods.

**Abstract** The size of simulations grids used for numerical models has increased by many orders of magnitude in the past years, and this trend is likely to continue. Efficient pixel-based geostatistical simulation algorithms have been developed, but for very large grids and complex spatial models, computational burden remains heavy. As cluster computers become widely available, using parallel strategies is a natural step for increasing the usable grid size and the complexity of the models. These strategies must take profit of the possibilities offered by machines with a large number of processors. On such machines, the bottleneck is often the communication time between processors. We present a strategy distributing grid nodes among all available processors while minimizing communication and latency times. It consists in centralizing the simulation on a master processor that calls other slave processors as if they were functions simulating one node every time. The key is to decouple the sending and the receiving operations to avoid synchronization. Centralization allows having a conflict management system ensuring that nodes being simulated simultaneously do not interfere in terms of neighborhood. The strategy is computationally efficient and is versatile enough to be applicable to all random path based simulations methods.

## 1. Introduction

The size of the simulation grids used for geological models (and more generally for spatial statistics) has increased by many orders of magnitude in the last years. This trend is likely to continue because the only way of modeling different scales together is to use high-resolution models. This is of utmost importance in applications such as hydrogeology, petroleum and mining, due to the critical influence of small scale heterogeneity on large scale processes (e.g. Mariethoz, et al., 2009).

Efficient pixel-based geostatistical simulation algorithms have been developed, but for very large grids and complex spatial models, the computational burden remains heavy. Furthermore, with increasingly sophisticated simulation techniques including complex spatial constraints, the computational cost for simulating one grid node has also raised. As multicore processors and clusters of computers become more and more available, using parallel strategies is necessary for increasing the usable grid size and hence allowing for models of higher complexity.

Parallel computers can be divided in two main categories: shared memory machines and distributed memory architectures. Shared memory machines have the advantage of ease and rapidity of the communications between the different computing units. Nevertheless, their price is extremely high and the total amount of

memory as well as the total number of processors are limited. Therefore, most of the time it is distributed memory machines (or clusters computers) that are used in the industry or in the academic world. As such machines do not have a common shared memory space, the processors have to communicate by sending and receiving messages. The communication time between processors can be important and is often the bottleneck in a program execution.

In this paper, we propose a parallelization strategy applicable in the context of sequential simulation methods and based on the distribution of the grid nodes among all available processors. The method minimizes communication and latency times and can be applied using shared or distributed memory architectures, or a combination of both. It consists in centralizing the simulation on a master processor that calls other slave processors as if they were functions simulating one node each time. The key is to decouple the sending and the receiving operations to avoid waiting for synchronization. Centralization allows having a conflict management system making sure that nodes being simulated simultaneously do not interfere in terms of neighborhood.

The strategy is computationally efficient and is versatile enough to be applicable to all random path based simulation methods. It is illustrated with an example using the Direct Sampling approach (Mariethoz, et al., submitted), which is a simulation algorithm using Multiple-Points (MP) statistics.

## 2. Parallelizing sequential simulations

Sequential simulation is a class of methods that is used to generate realizations of a random field (Deutsch and Journel, 1992; Caers, 2005). The general principle of the method is to discretize the random field on a grid and to draw successively (sequentially) for each node $\mathbf{x}$ of the grid an outcome of the random variable $Z$ in a local cumulative conditional density function (ccdf). This local ccdf is conditional to the previously simulated nodes and to local data if those are available. Usually, a truncation is made and the ccdf is computed only from the values located in a neighborhood $N(\mathbf{x})$ of limited extension.

The local ccdf is determined using a spatial model, termed $m$, that describes the spatial structure of the random field. Grid nodes are often simulated in a random order, but alternative simulation paths can also be used (Pickard, 1980; Daly, 2004).

Depending on the sequential simulation technique used, $m$ can be for example one or a set of variograms in the case of SGS / SIS simulations (Isaaks, 1984; Journel and Alabert, 1990), plus a lithotype rule for plurigaussian simulations (Le

Loc'h and Galli, 1994; Armstrong, et al., 2003), a training image or its associated data events catalogue for MP simulations (Strebelle, 2002; Zhang, et al., 2006; Arpat and Caers, 2007; Straubhaar, et al., 2008; Mariethoz, et al., submitted), or a set of transition probabilities (Carle and Fogg, 1997).

Each sequential simulation method has its own way of computing the value $z(m, N(\mathbf{x}))$ that will sequentially be attributed to each node $\mathbf{x}$. Nodes are simulated in an order defined by a random path initialized at the beginning of the simulation. Once a value has been attributed to a node, this node becomes conditioning for the nodes that come later in the simulation process, i.e. it will be included in the ensemble $N(\mathbf{x})$ for the next nodes to simulate. This is the reason why these simulation techniques are termed sequential.

Parallelization of such simulations is possible at three levels. The realization level is the easiest to parallelize. It consists in having each realization of a Monte-Carlo analysis computed by a different processor. As every realization is, by definition, independent of the others, no communications between processors are needed. The maximum number of processors that can be used with this strategy is equal to the desired number of realizations. This strategy is widely used (e.g. Mariethoz, et al., 2009) and will not be discussed further in this paper.

Parallelizing a simulation at the path level means to divide the grid in zones and to attribute a different zone to each processor. By now, this strategy has been implemented by simulating groups of grid nodes at the same time (Vargas, et al., 2007).

The third level of parallelization is the node level. The simulation of each single node is parallelized. For example, the inversion of a large kriging matrix for SGS or the search for a data event in the multiple-points data events catalogue can be shared among many processors (Straubhaar*, et al.*, 2008). Speculative parallel computing can also be applied in the context of simulated annealing (Ortiz and Peredo, 2008). In all of these cases, the efficiency of the parallelization is limited when a large number of processors are available, because the size of the problem to solve for each individual processor becomes small compared to the communications time between processors.

These different strategies are not mutually exclusive. For example, the path can be distributed among different parallel machines, who themselves distribute the simulation of their individual nodes on local processors.

This paper focuses on parallelization strategies at the path level. The sequential character of the simulation process is a challenge for these strategies because of the dependence of the value of $z(\mathbf{x})$ with all previously simulated nodes. Another issue

is that the time taken to simulate a node is not necessarily uniform, depending on the simulation method. Some nodes can be simulated much slower than others, which have for example a neighborhood less compatible with the spatial structure model *m*. In some cases, the number of neighbors can be different from one node to another, incurring variations in computational load. This problem becomes more acute when the simulation algorithm is run on heterogeneous architectures mixing processors of different performance.

Parallelizing the random path will inevitably lead to conflicts when a node has to be simulated by a processor while nodes of its neighborhood are being simulated by other processors. Moreover, certain algorithms, such as the Gibbs sampler (Geman and Geman, 1984) or the syn-processing (Mariethoz, et al., submitted), require to re-simulate nodes that don't match certain conditions. This complicates the problem as it leads to changes in the simulation path, making it impossible to define in advance a conflicts-free path. In certain cases, the simulation method can be adapted to be less sensitive to these conflicts (Dimitrakopoulos and Luo, 2004). But our goal is to find a general strategy that is applicable to all random path based simulation methods without generating conflicts.

## 3. Nodes distribution

The solution proposed in this paper is to have one processor, the master, managing the path, the search for neighbors and the conflicts, while all other processors, the slaves, devote their calculation power to the simulation itself. If $n_{CPU}$ processors are available, the processor 0 is the master and processors 1 to $n_{CPU}$-1 are the slaves. The most obvious strategy would be to group $n_{CPU}$-1 nodes and distribute them among slave processors, a strategy adopted by Vargas, et al. (2007). Unfortunately, this strategy is not efficient with a large number of processors because it involves that all slaves must have returned their result to the master before the next group of nodes is sent to slaves for simulation. If one of the slaves uses more time than the others to simulate his node, all processors have to wait for it to finish. Moreover, the master does not perform any calculations while slaves are working, and the slaves also have to wait until the master has finished updating the simulation with the received group of nodes and has defined the neighborhoods for the next group.

Instead of groups, we propose that the master sends sequentially one node to each slave, and then waits for a result coming from any slave processor. Once this result is obtained, the master includes it in the simulated grid, finds the

neighborhood of the next node and sends it to the same slave processor from which the result just came from. Then it waits again for a result coming from any slave processor. By avoiding synchronization, this strategy ensures that a minimum number of processors are waiting. The master practically does not wait when there are a lot of slaves. Moreover, while the master works on defining neighborhoods and attributing nodes to slave processors, all slaves except one are working. The slaves do not wait for each other as they perform their workload independently. The time devoted to communications is minimized because while the master sends or receives information from a slave, all other slaves carry on their work undisturbed.

The procedure needs a starting routine that sends the first $n_{CPU}$-1 nodes on each slave CPU without receiving any result, and another ending routine that receives the last $n_{CPU}$-1 nodes from each slave CPU.

This way of centralizing the simulation on a master node has many implications. First, this strategy is impossible to apply on machines having a single processor. Second, it is not efficient if the total number of processors is not large enough. For example, using 2 processors is absurd as only one of them (the slave) is performing the simulation while the other one (the master) waits. But with an increasing number of processors, the communication time of each slave decreases, while it stays constant for the master, which can afford it because it practically does not perform any computation.

Another implication of the centralization is that it is easy to efficiently implement algorithms that come back on previously simulated nodes, such as for example the Gibbs sampler or the syn-processing. If changes must happen in the simulation path or in previously simulated nodes, the master node alone has to manage it and no additional communications have to take place.

A disturbing consequence of the strategy is that it does not allow reproducibility of the resulting simulations, although all of them will be consistent with the spatial law *m*. The random order in which nodes are sent to slaves and are returned in the simulation grid depends on the local characteristics and on the workload of each processor. Therefore, the order in which the nodes are simulated is continuously altered during the simulation process, which is equivalent to using a different random simulation path. As the simulated path is random anyway, this does not affect the faithfulness to the spatial model. But reproducibility of the simulations is compromised because the simulation path is not reproducible (it depends on complex hardware and software interactions that are difficult to control). Different runs of the same code yield different simulations, even if the random seed is identical.

It is also important to note that this strategy is inefficient when using an unilateral path (Daly, 2004) because the amount of conflicts generated would make it inefficient. It is also not directly applicable to techniques other than sequential simulations, such as boolean simulation (Deutsch and Tran, 2002; Lantuéjoul, 2002) or turning bands (Matheron, 1973), for which other parallelization methods have been developed (Armstrong and Marciano, 1997; Kerry and Hawick, 1998; Ingam and Cornford, 2008).

## 4. Conflicts management

Let $U_{sim}$ be the ensemble of all nodes currently simulated by all slave processors (ensemble containing $n_{CPU}$-1 elements). When a new node $\mathbf{x}$ has to be simulated, conflicts arise when $N(\mathbf{x}) \cap U_{sim} \neq \emptyset$, i.e. when at least one node currently simulated by any slave processor belongs to the neighborhood of node $\mathbf{x}$.

When a conflict arises, one can consider three ways of dealing with it. The first one is to ignore the conflict. This option is worth considering if the simulation grid is large and the number of processors reasonable, thus a limited number of conflicts will occur and it can be assumed that the resulting simulation will not be significantly biased. The second option is to try simulating another node $\mathbf{x}$ until there is no more conflict. Most of the time, the conflict will disappear after a certain number of tries, but there are cases where there is no suitable location in the entire simulation grid that generates no conflict. The conflict is then severe and only the third option remains: waiting until some of the nodes creating conflicts are simulated, and then try again.

The strategy adopted here combines the second and the third options. The idea is to try the second option $n_{tries}$ times. After this maximum number of tries, one can consider that there is no conflict-free location in the entire simulation grid. At this point, the third option is used: waiting for slaves to return their simulated nodes.

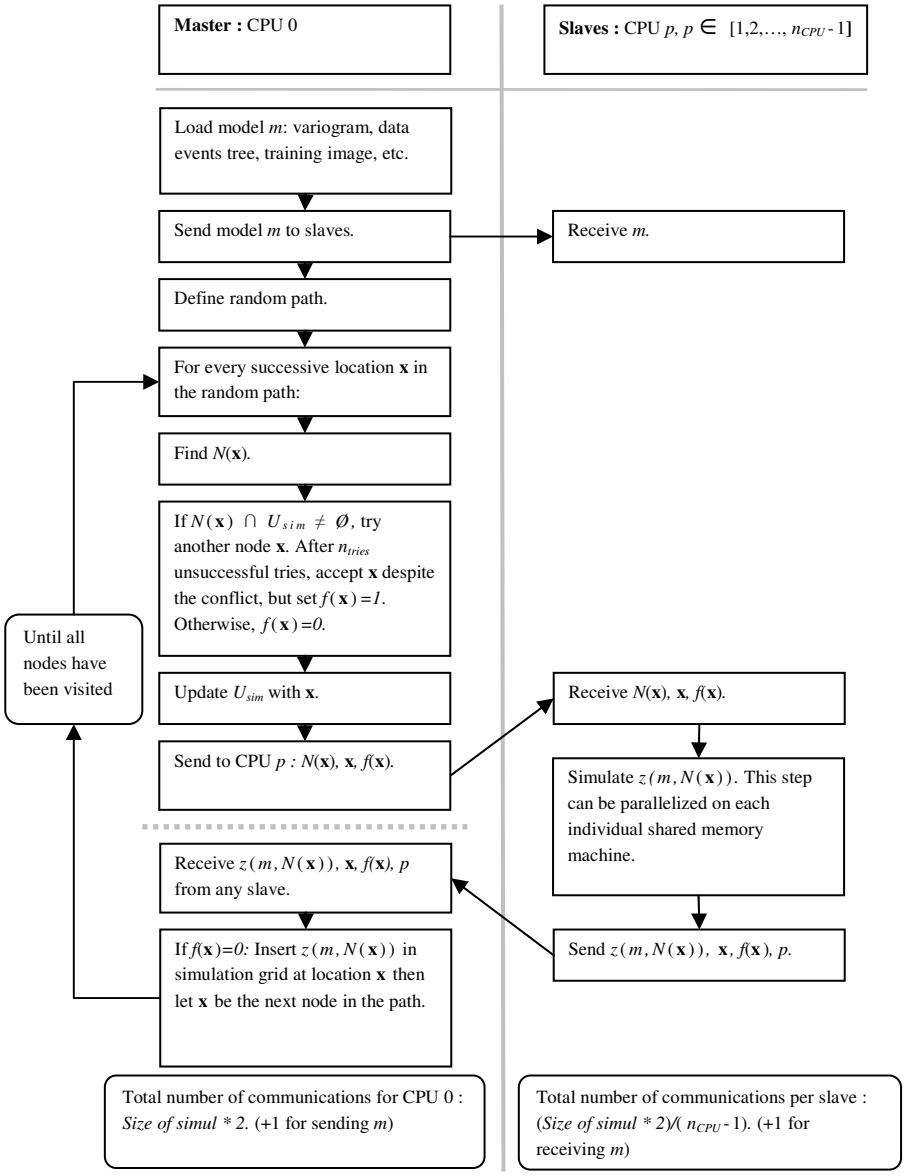



**Figure 1** Sketch of the proposed parallelization strategy. Each column describes schematically the series of instructions executed by, respectively, the master and the slaves. The meaning of each variable is described in the text. The dashed horizontal line represents the lapse when the master waits for a result coming from any slave CPU.

In the context of node distribution, simply waiting for a returning node is not trivial. The procedure involves that for each returned node, another node has to be sent. The only way to receive a node is to send another node, even if it presents a conflict. The solution is to mark this node as "false", by setting the flag function $f(\mathbf{x})$ to 1 to signal that this node is causing a severe conflict, and send it to any slave processor. When a value for this "false" node is returned by the slave processor, it is discarded and simulated again with a different, conflict-free neighborhood. There are no more conflicts because in the meantime, the neighbors causing conflicts have been simulated. Figure 1 schematically describes the entire strategy.

Note that nodes distribution and conflicts management involve that the random path defined at the initial stage of the simulation is continuously changed depending on neighborhood interactions and individual computational load of each processor. The path updating is done by the master.

## 5. Performance tests

The parallelization strategy described above has been tested on the Direct Sampling algorithm (Mariethoz, et al., submitted), a recent implementation of Multiple-points simulation (Guardiano and Srivastava, 1993; Strebelle, 2002; Hu and Chugunova, 2008). For each node **x** in the simulation grid, the algorithm scans a training image (TI) representing what the simulated field should look like independently of the data. As soon as a pattern matching the neighborhood of **x** in the simulation is found in the TI, the value of the central node of this pattern is assigned at location **x** of the simulation grid.

The performance tests consist in generating, with a varying number of processors, one unconditional realization on a relatively small grid (100 by 100 by 5 nodes). The neighborhoods are large (search radius size ranging from 30 nodes at the beginning of the simulation to 3 nodes at the end), thus likely to generate conflicts. With 14 million nodes, the training image is much larger than the simulation, and therefore scanning it for each simulated node represents a steep cost in terms of CPU time. The MPI libraries were used as a parallel interface. The machine available for the tests is a cluster of 56 AMD Opteron processors, grouped in 14 machines of 4 processors each, with InfiniBand interconnect.

Simulations were performed using a total number of processors ranging from 2 to 54. Two series of runs were performed. The first series addresses conflicts as described above (combination of the second and the third options for dealing with

conflicts), with $n_{tries}$ set to 1'000. The second series ignores conflicts. Comparing both series allows evaluating the time spent for conflicts management.

Figure 2 shows the decrease of CPU time as the number of slave processors becomes larger. The measure of CPU time is the user time of the slowest processor of each run (master and slaves), in seconds. Both series of runs are displayed. The series ignoring conflicts is expected to be faster than the one accounting for it, but such is not always the case. We explain this by fluctuations in the simulation time from one simulation to another. One should keep in mind that the simulations are not reproducible. Depending on the simulation path, more or less probable patterns may be generated, causing a different fraction of the TI to be scanned for each simulated node (this fraction has a large influence on the simulation time). It is an inherent characteristic of the parallelization strategy that the path is different for each realization. Therefore, computation times are subject to fluctuations that cannot be controlled.

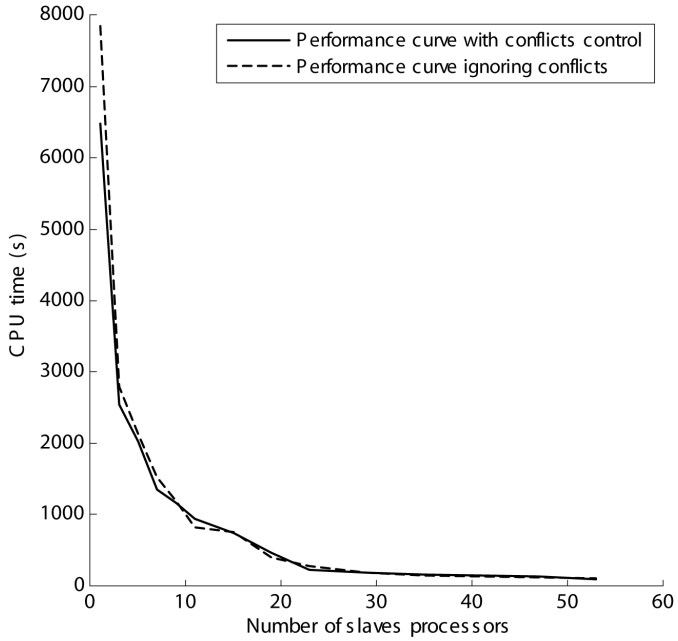



**Figure 2** Performance curve on a grid of 50'000 nodes (100 by 100 by 5), with and without conflicts control. Conflicts management consists in first trying $n_{tries}$ times another location **x** to simulate, and if no suitable location is found, wait for the conflict to disappear. Conflicts management does not significantly influence the performance.

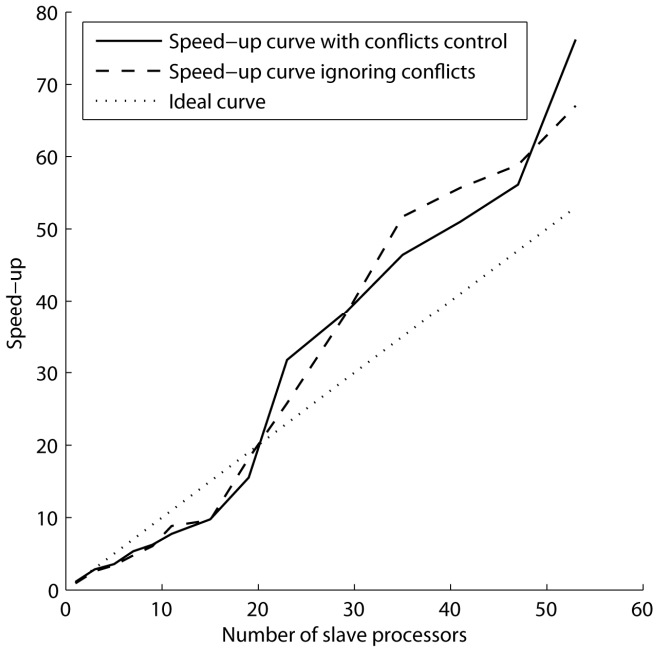



**Figure 3** speed-up curves on a grid of 50'000 nodes (100 by 100 by 5), with and without conflicts control.

A common performance test for parallel algorithms is to compute the speed-up function, which is the ratio between the time taken by the algorithm to perform some computations in a serial way and in parallel. It is defined as:

$$sp(n_{slaves}) = \frac{t_{serial}}{t(n_{slaves})} = \frac{t_{serial}}{t(n_{CPU} - 1)}, \tag{1}$$

where $t$ is the CPU time needed to solve a given problem with a certain number of processors. In the best case, the speed-up should follow the ideal slope $sp(n_{slaves})=n_{slaves}$ (Amdahl, 1967). Figure 3 shows the speed-up functions for both series of runs. In our case, the number of slave processors is considered instead of the total number of processors, because using a single processor is impossible. The reference time $t_{serial}$ was obtained by running the serial version of the Direct Sampling code with the same parameters, on the same machine.

It is not straightforward to explain why the speed-up function (Figure 3) exceeds the ideal curve when more than 16 processors are used. Moreover, the serial version is slower (7'240 seconds) than using only one slave (6'483 seconds), which is not logical because the overhead caused by the communications should degrade the performance. One possible reason could be related to the optimized use of the cache memory in the parallel version. Another reason could be that the reference time was computed using a different code (parallel and serial versions are fundamentally different, although it is the same simulation algorithm), which could induce a bias. Fine analysis of both codes may give clues on the cause for this speed-up increase. In any case, the speed-up in this example is good when the number of processors is large, which is a sign that the method is efficient.

For a parallel algorithm to be efficient, the workload must be balanced between all available processors. If large discrepancies are observed between the times of the different CPUs, it means that the workload is not well shared. Figure 4 displays the load balancing of each processor in 9 of the performance tests with conflicts control. These curves show that there are no large CPU time differences between the processors of a same run. Moreover, when the number of processors becomes larger, fewer discrepancies appear. This result indicates that there are no major parallelization bottlenecks slowing down the entire simulation. Here again, one can observe that the CPU time is very well balanced when more than 16 processors are used.

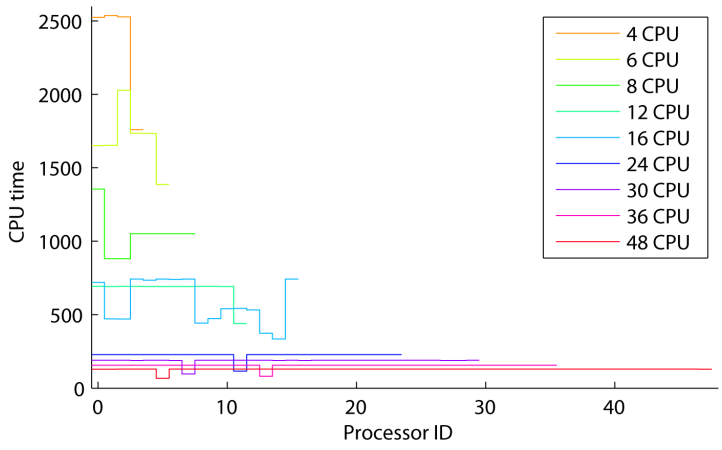



**Figure 4** Load balancing on a grid of 50'000 nodes, with conflicts control. The CPU time of each processor is represented. In an ideal case, all processors would consume the same amount of CPU time.

In these performance tests, conflicts management does not significantly influence performance. Table 1 shows that even if the total number of tries (second option: the total number of times that the master node tried to solve a conflict by trying to simulate another node **x**) is very large, sometimes larger than the number of nodes in the simulation, it does not have a major influence on the global simulation time. This is because the computational cost of trying another node is small compared to the cost of simulating a node.

In case of severe conflicts (when no conflict-free location can be found in the entire simulation grid), the master has to send to a slave a node marked as "false", in order to wait other slaves return their result, causing the severe conflict to disappear. This way of solving severe conflicts (third option) is much more time-consuming than the second option, but fortunately it happens much less often (compare values of rows "total tries" and "total wait" in table 1).

| $n_{CPU}$ | **2** | **4** | **6** | **8** | **10** | **12** | **16** |
|---|---|---|---|---|---|---|---|
| **Max time (s)** | 6'483 | 2'537 | 2'028 | 1'355 | 1'158 | 934 | 743 |
| **Total tries** | 0 | 313 | 877 | 1'491 | 54'055 | 2'831 | 129'130 |
| **Total wait** | 0 | 0 | 0 | 1 | 54 | 2 | 129 |

| $n_{CPU}$ | **20** | **24** | **30** | **36** | **42** | **48** | **54** |
|---|---|---|---|---|---|---|---|
| **Max time (s)** | 467 | 228 | 189 | 156 | 142 | 129 | 95 |
| **Total tries** | 31'032 | 44'045 | 86'088 | 147'268 | 251'421 | 746'746 | 272'272 |
| **Total wait** | 31 | 44 | 86 | 147 | 251 | 746 | 272 |

**Table 1** Numerical results of the performance tests with conflicts management. Max time is the CPU time of the slowest processor in the run. Total tries is the total number of times that another node was tried in case of conflicts (second option). Total wait is the total number of times that a "false" node had to be simulated (third option) because the second option failed. Logically, Total wait is 1'000 times smaller than Total tries because $n_{tries}$ was set to 1'000.

The computational burden associated to conflicts management only depends on the size of the simulation grid and on the number of processors. If the cost of simulating one node is small, the time devoted to conflicts management may be comparatively larger. Performance is most affected by conflicts management in cases where the simulation grid is small, the number of processors is large and the

cost per simulated node is low. Nevertheless, in these cases realizations should be quickly obtained with a serial algorithm, and parallelization is not needed.

If the number of processors is very large, the workload of the master node increases because of the extra cost of managing conflicts between all processors. Although we did not observe it during numerical tests with up to 54 processors, there should be a number of slaves where the master becomes the bottleneck because there are simply too many slaves to manage. This could affect performance on very large problems. A solution would be to split the work of the master in two levels by having not one but several master processors, and one super-master managing the masters. Following this idea, one could imagine systems with three or more levels of management.

## 6. Conclusion

Among the three simulation parallelization levels (realization, path and node), the path level allows using a large number of processors without losing efficiency. It is applicable to all random path based geostatistical simulation methods, even if a small number of simulations have to be generated (contrarily to the realization level) In this paper, we propose a strategy that takes full profit of this parallelization level.

The overall performance of the strategy is good according to the speed-up criterion. This confirms that the communication and latency times are minimal and that most of the computational effort of all processors is devoted to the simulation and not to the parallelization overload.

The issue of overlapping neighborhoods arising in the context of nodes distribution is addressed in a two-step process that tries to quickly solve the conflict by trying to find a conflict-free location to simulate, and in case of failure waits for the conflict to disappear. Conflicts management ensures that the quality of the realizations obtained with a parallelized simulation algorithm is the same as with the serial algorithm. In terms of performance, tests showed that conflicts management does not have a major influence.

The proposed parallelization strategy is flexible in many ways. It allows accelerating the stochastic image generation process independently of the simulation technique used. It can also accommodate different simulation techniques as well as various computer architectures such as grid computing or heterogeneous clusters.

The main drawback of the method is that it does not allow reproducibility of the resulting simulations.

## 7. References


Amdahl, G. (1967), *Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities*, paper presented at AFIPS Conference Proceedings, Thompson Books, Washington D. C.

Armstrong, M., and Marciano, R. (1997), *Massively parallel strategies for local spatial interpolation* Computers & Geosciences, 23, 8, 859-867.

Armstrong, M., Galli, A. G., Loc'h, G. L., Geoffroy, F., and Eschard, R. (2003), *Plurigaussian Simulations in Geosciences*, Springer.

Arpat, B., and Caers, J. (2007), *Conditional Simulations with Patterns*, Mathematical Geology, 39, 2, 177-203.

Caers, J. (2005), *Petroleum Geostatistics*, Society of Petroleum Engineers.

Carle, S. F., and Fogg, G. E. (1997), *Modeling spatial variability with one and multi-dimensional continuous Markov chains*, Mathematical Geology, 7, 29, 891-918.

Daly, C. (2004), *Higher order models using entropy, Markov random fields and sequential simulation*, paper presented at Geostatistics Banff 2004, Kluwer Academic Publisher, Banff, Alberta.

Deutsch, C., and Journel, A. (1992), *GSLIB: Geostatistical Software Library*, Oxford Univ. Press.

Deutsch, C., and Tran, T. (2002), *FLUVSIM: a program for object-based stochastic modeling of fluvial depositional systems*, Computers & Geosciences, 2002, 28, 525-535.

Dimitrakopoulos, R., and Luo, X. (2004), *Generalized Sequential Gaussian Simulation on Group Size v and Screen-Effect Approximations for Large Field Simulations*, Mathematical Geology, 36, 5, 567-591.

Geman, S., and Geman, D. (1984), *Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images*, IEE Trans. Pattern Anal. and Mach Intel., 6, 6, 721-741.

Guardiano, F., and Srivastava, M. (1993), *Multivariate geostatistics: Beyond bivariate moments*, in *Geostatistics-Troia*, pp. 133-144, Kluwier Academic.

Hu, L., and Chugunova, T. (2008), *Multiple-Point Geostatistics for Modeling Subsurface Heterogeneity: a Comprehensive Review*, Water Resour. Res., 44, W11413.

Ingam, B., and Cornford, D. (2008), *Parallel geostatistics for sparse and dense datasets*, paper presented at geoENV 2008, Southampton, 8-10 September, 2008.

Isaaks, E. (1984), *Indicator simulation: Application to the simulation of a high grade uranium mineralization*, in *Geostatistics for Natural Resources Characterization, Part 2*, pp. 1057-1069, D. Reidel Publishing Company.

Journel, A., and Alabert, F. (1990), *New Method for Reservoir Mapping*, Journal of Petroleum Technology, 42, SPE paper 20781, 212-218.

Kerry, K., and Hawick, K. (1998), *Kriging Interpolation on High-Performance computers*, paper presented at the International Conference and Exhibition on High-Performance Computing and Networking, Amsterdam, the Netherlands, April 21-23, 1998.

Lantuéjoul, C. (2002), *Geostatistical simulation. Models and algorithms.*, Springer.

Le Loc'h, G., and Galli, A. G. (1994), *Improvement in the truncated Gaussian method: combining several Gaussian functions*, paper presented at Ecmor 4, 4th European Conference on the Mathematics of Oil Recovery, Roros, Norway.

Mariethoz, G., Renard, P., Cornaton, F., and Jaquet, O. (2009), *Truncated plurigaussian simulations to characterize aquifer heterogeneity*, Ground Water, 47, 1, 13-24.

Mariethoz, G., Renard, P., and Straubhaar, J. (submitted), *The direct sampling method to perform multiple-points simulations*, Water Resour. Res.

Matheron, G. (1973), *The Intrinsic Random Functions and their applications*, Advances in Applied Probability, 5, 439-468.

Ortiz, J. M., and Peredo, O. (2008), *Multiple Point Geostatistical Simulation with Simulated Annealing: Implementation using Speculative Parallel Computing*, paper presented at geoENV 2008, 7th International Conference on Geostatistics for Environmental Applications, 8-10 Sept. 2008, Southampton.

Pickard, D. (1980), *Unilateral Markov fields*, Advances in Applied Probability, 12, 655-671.

Straubhaar, J., Walgenwitz, A., Renard, P., and Froidevaux, R. (2008), *Optimization issues in 3D multipoint statistics simulation*, paper presented at Geostats 2008, 1-5 Dec. 2008, Santiago, Chile.

Strebelle, S. (2002), *Conditional simulation of complex geological structures using multiple point statistics.*, Mathematical Geology, 34, 1, 1-22.

Vargas, H., Caetano, H., and Filipe, M. (2007), *Parallelization of sequential Simulation Procedures*, paper presented at Petroleum Geostatistics, EAGE, Cascais, Portugal.

Zhang, T., Switzer, P., and Journel, A. (2006), *Filter-Based Classification of Training Image Patterns for Spatial Simulation*, Mathematical Geology, 38, 1, 63-80.

# Chapter 6

# Integrating collocated auxiliary parameters in geostatistical simulations using joint probability distributions and probability aggregation *

*[About Gaussian models:] "In no field of empirical inquiry has so massive and sophisticated a statistical machinery been used with such indifferent results."*

*Wassily Leontief*

* This chapter is based in the paper:
Mariethoz, G., P. Renard, R. Froidevaux. Integrating collocated auxiliary parameters in geostatistical simulations using joint probability distributions and probability aggregation, Water Resour. Res., 45, W08421, doi:10.1029/2008WR007408.

**Abstract** We propose a new cosimulation algorithm for simulating a primary attribute using one or several secondary attributes known exhaustively on the domain. This problem is frequently encountered in surface and groundwater hydrology when a variable of interest is measured only at a discrete number of locations and when a secondary variable is mapped by indirect techniques such as geophysics or remote sensing. In the proposed approach, the correlation between the two variables is modeled by a joint probability distribution function. A technique to construct such relations using latent variables and physical laws is proposed when field data are insufficient. The simulation algorithm proceeds sequentially. At each node of the grid, two conditional probability distribution functions (cpdf) are inferred. The first is inferred in a classical way from the neighboring data of the main attribute and a model of its spatial variability. The second is inferred directly from the joint probability distribution function of the two attributes and the value of the secondary attribute at the location to be simulated. The two distribution functions are combined by probability aggregation to obtain the local cpdf from which a value is randomly drawn. Various examples using synthetic and remote sensing data demonstrates that the method is more accurate than the classical collocated cosimulation technique when a complex relation relates the two attributes.

## 1. Introduction

There are numerous situations in surface and groundwater hydrology in which an attribute of interest is measured at a discrete number of locations and needs to be mapped accounting for exhaustive secondary information. For example this is encountered 1) when interpolating ground based measurements of precipitation using a digital elevation model (Goovaerts, 2000) or radar observations (Creutin*, et al.*, 1988; Haberlandt, 2007) as secondary information; 2) when estimating soil surface moisture and surface roughness from satellite images and ground measurements (Makkeasorn*, et al.*, 2006); 3) when characterizing the heterogeneity of an aquifer from local hydraulic conductivity data and an exhaustive geophysical survey (Cassiani*, et al.*, 1998); 4) when mapping groundwater recharge using local estimates and an exhaustive evaporation map produced from remote sensing data (Brunner*, et al.*, 2004); or 5) when mapping large scale groundwater contamination (such as Arsenic) using again local measurements in boreholes and exhaustive maps of geology and surface soil properties (Winkel*, et al.*, 2008).

The simplest way to solve this type of problem is to model the correlation between the primary and secondary attributes using a statistical regression (it does not need to be linear) and to rescale the exhaustive map. This is, for example, the procedure that was used by Brunner et al. (2004) for recharge estimate and by

Winkel et al (2008) to map Arsenic concentration. However such a method does not honour the data (the interpolated value at the location of a ground based measurement may be different from the measurement itself) and does not account for the possibly known covariance structure of the primary variable. More flexible approaches are available in the framework of multivariate geostatistics (Chilès and Delfiner, 1999; Wackernagel, 2003). Such methods were reviewed and compared in the field of hydrology for example by Ahmed and de Marsily (1987) or Goovaerts (2000). They include a set of estimation techniques (providing the most likely value at any location) such as cokriging (Matheron, 1971), kriging with external drift (Delhomme*, et al.*, 1981), collocated cokriging (Xu*, et al.*, 1992), and stochastic simulation techniques (allowing to generate a set of equally probable maps) such as cosimulation (Gomez-Hernandez and Journel, 1993), collocated cosimulation, cosimulation with external drift, etc. These techniques are based on two ingredients: a model of cross-covariance that relates the variability of the first variable at a given location with the variability of the secondary variable at another location, and the assumption that the relation between the two variables is essentially linear and can therefore be modeled in a multi-Gaussian framework. Furthermore, it is often found in practice that the secondary data located precisely at the location that needs to be estimated (or simulated) has a much larger impact on the estimation (or simulation) than the data located aside. This is why when an exhaustive map of secondary information is available, a common simplification is to write a cokriging system that accounts only for the secondary variable at the location to be estimated (or simulated) and not at the neighboring nodes. These are the so-called collocated cokriging or cosimulation techniques (Xu*, et al.*, 1992). It is an approximation that has the advantage to be much faster than the full cokriging but it is not optimal for all situations and can even be useless in certain cases. For example, when solving an inverse problem in a cokriging framework, the cross-covariance between heads and transmissivity under the uniform flow assumption shows that a head measurement has no impact (cross-covariance equal to zero) on the estimation of the transmissivity at the location of the head measurement (Dagan, 1989).

In this paper, we do not consider the inverse problem. Our aim is to propose a new method for generating an ensemble of stochastic simulations of a primary attribute using a discrete number of local measurements of the primary attribute and an exhaustive map of the secondary attribute in a collocated cosimulation framework. The specificity of our approach is to consider complex relations between the two variables and to use a probability aggregation technique.

Indeed, most papers presenting applications of geostatistical techniques (cokriging or cosimulations) to this problem assume a multiGaussian framework and a linear relationship (Figures. 1a and 1d) between the two attributes. When the relation is not linear but can be modeled by an analytical function (e.g. Figure 1b), one can use a transformation of variable to linearize the relation. For example this is the case when estimating the log hydraulic conductivity using geoelectrical surveys (Cassiani*, et al.*, 1998; El Idrisy and De Smedt, 2007; Slater and Lesmes, 2002; Soupios*, et al.*, 2007). However, the assumptions of multiGaussianity and linear correlation between the variables (or their transforms) is too restrictive. It is an oversimplified description of complex physical processes and therefore it does not hold in many real-case applications. For example, we observed in a karstic coastal aquifer in Oman (Alcolea*, et al.*, 2009) that karstic conduits had a high hydraulic conductivity and a low electrical resistivity if they were fully saturated with sea water. But clayey deposits with very low hydraulic conductivity (that were also saturated with seawater) had a low electrical resistivity too. Higher electrical resistivity were associated to intermediate conductivity. A low electrical resistivity could therefore indicate either a high or a low hydraulic conductivity, but not an intermediate conductivity (Figure 1c). Such a relation was not only non linear (see Figure 1b for an example of nonlinear relation), it was essentially non-injective or multi-valued (Figure 1c). We think that such situations are more frequent than usually admitted and that there are many cases in which there may be several possible and distinct values (or modes) of the primary variable for a given value of the secondary variable. This type of non-injective relations cannot be modelled by an analytical function of the secondary variable only nor by a simple change of variable. In addition, the possible variability of the primary variable knowing the secondary variable can vary as a function of the value of the secondary variable (heteroscedasticity) as shown in Figure 1e.

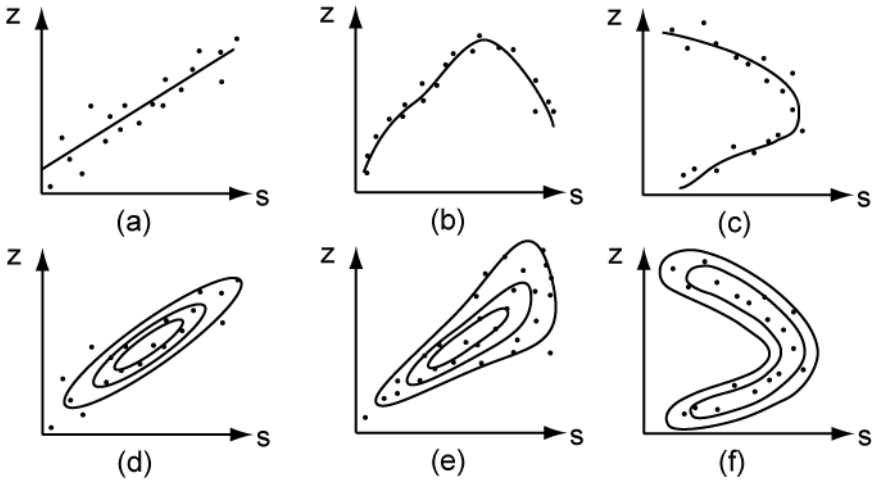



**Figure 1** Schematical representation of the different types of possible relationships between a primary and secondary variable. *s* represents the secondary variable, *z* the primary one. (a) linear relationship modeled by a regression line, (b) non linear relationship modeled by an analytical injective function relating $z = f(s)$, (c) non-injective (or multi-valued) relationship between *s* and *z* which cannot be modeled by a function $z = f(s)$, (d) linear relationship modeled by a biGaussian probability distribution function, (e) linear relationship showing a variation of the uncertainty around the regression line which cannot be modeled by a biGaussian pdf, (f) non-injective relation modeled by a joint probability distribution.

A possibility to model the relation between the two variables in a completely general and statistical manner is to use a bivariate probability distribution function (pdf) $f(z,s)$. This joint pdf can either be expressed analytically and parameterized by a type of distribution and its means, variances, correlation coefficient, etc., or it can be expressed as a fully non-parametric joint distribution (i.e. provided as a numerical matrix of probability values such as a bivariate histogram).

To our knowledge, only a few techniques allow the use of complex bivariate models. The bilinear coregionalization model (Wackernagel, 2003) is a step towards integrating more complex relationships as it allows for non even covariance functions in the multiGaussian framework. The so-called cloud transform technique (Bashore*, et al.*, 1994; Kolbjørnsen and Abrahamsen, 2004) and the stepwise conditional transformation technique (Leuangthong and Deutsch, 2003), are closely related in the sense that they both impose a given bivariate correlation by

performing appropriate normal score transformations of the variables. These methods rely also on a non-parametric description of the bivariate distribution between the main and secondary attributes thus allowing to accommodate situations in which the marginal distributions are multimodal and the relationship between the attributes is non-linear. Simulated annealing (Deutsch, 1992) is another technique that could be used to impose various kinds of constraints, including non-parametric correlations (Caers, 2001). Thanks to its flexibility, it has been applied to various practical cases (Dafflon*, et al.*, 2008), even if it presents shortcomings such as a high CPU cost and a high parameterization of the cooling schedule.

In all the methods cited above, a single conditional distribution function of the main attribute is estimated directly for each simulated grid node. Here, we split the problem and estimate two separate probability distribution functions. The first one is estimated considering the primary variable only and a model of its spatial variability. In order to respect the nature of the first variable, that may or may not be multiGaussian, any suitable technique can be used (e.g. multiGaussian kriging (Emery, 2005; Goovaerts*, et al.*, 2005; Verly, 1993), direct sequential simulation (Soares, 2001), indicator kriging (Journel and Alabert, 1990; Journel and Isaaks, 1984), multiple points statistics (Hu and Chugunova, 2008), etc.). The second one is the conditional distribution of the primary attribute computed from the bivariate distribution knowing the value of the secondary attribute. These two distributions are then combined into a single one using the concept of probability conjunction (Tarantola, 2005), which is a particular case of the theory of Bordley (1982) used in management science for aggregating expert's opinions. A similar approach was used by Ortiz and Deutsch (2004) to update the indicator kriging probability with multiple-points statistics.

To apply the technique proposed in this paper, one needs first to model the bivariate distribution describing the relation between the two attributes. This is a prerequisite of the method. Because there are many statistical techniques available to do so, we will not focus on that aspect within the paper but just give some directions to help the user. Assuming that enough couples of values relating the primary and the secondary variable are available, one can statistically infer the parameters of any analytical joint distribution, or use techniques such as kernel smoothing to build a non parametric distribution (Epanechnikov, 1969; Kolbjørnsen and Abrahamsen, 2004). A thorough review of the techniques for estimating non-parametric density functions can be found in Izenman (1991). Even if co-located data are available, the problem of estimating a statistical relationship between geophysical and hydrologic properties is not trivial. Several caveats remain, such as of scale issues (Moysey and

Knight, 2004), local variations in the relationships and artifacts caused by inversion techniques (Day-Lewis and Lane, 2004). Some authors have addressed these issues and proposed sophisticated and efficient techniques for estimating such relationships based on rock physics relationships and Monte-Carlo simulations (Moysey, *et al.*, 2005; Mukerji, *et al.*, 2001). When there is not enough data, we propose an alternative approach that consists in building the bivariate distribution from known physical laws and latent variables that indirectly relate the attributes of interest.

The paper is structured as follows. The first part proposes a method for building a physically based joint probability distribution function when there is a lack of direct data. The second part describes the proposed simulation algorithm. The third part illustrates its application on a fully synthetic example. The fourth part presents an application of the method on a more realistic example based on satellite images. That case is used because, even if it has no direct application, it is analogous to real problems such as those described above and, more importantly, it is one of the rare situation that allows to test the accuracy of the method with real data showing a complex bivariate relation. This example is also used to show how to determine the optimal parameters of the simulation algorithm. Last comes a discussion of the overall method.

## 2. Inferring the joint distribution from physical laws and latent variables

Denote:

| | |
|---|---|
| $\mathbf{x}$ | vector describing a location. |
| $Z(\mathbf{x})$ | the attribute of main interest. |
| $S(\mathbf{x})$ | the co-located attribute. |
| $z(\mathbf{x}_i),\ i=[1\dots N]$ | available conditioning data for the main attribute at location $\mathbf{x}_i$. |
| $f(z,s)$ | bivariate joint probability density function. |

Note that upper cases $S$ and $Z$ represents random functions, while lower cases $s$ and $z$ represents actual values of these random functions.

When there is not enough data available to statistically infer the joint distribution $f(z,s)$, one can build a realistic distribution from physical laws and latent variables. Often the variables of interest ($z$ and $s$) are indirectly related to one (or several) underlying attribute $t$ through physical laws: $z=\alpha(t)$ and $s=\beta(t)$. The underlying attribute $t$ is the latent variable (Bollen, 1989). Often there is sufficient data to

estimate or assume the univariate pdf $f(t)$. To build the joint pdf $f(z,s)$, one can randomly sample a large number of values of $t$ in $f(t)$, compute the corresponding values of $z$ and $s$ for each value of $t$ and obtain a large number of couples $[z(t),s(t)]$ for estimating the joint pdf.

The following example illustrates this method. The primary attribute is the hydraulic conductivity and the secondary attribute is the electrical resistivity. To build a relation between these two attributes, we use porosity as the latent variable. The data used in this example originates from core samples of USM (Untere Süsswasser Molasse) level of the Swiss Molasse formation, in which four main facies have been identified. The overall composition of the USM is mainly marl and sandstone. Field data are described in Mariethoz, *et al.* (2009).

The first physical law is the Hagen-Poiseuille equation (1) relating hydraulic conductivity $K$ [m/s] to porosity $\phi$ [-]:

$$K = \frac{\phi^3 \rho g}{b A_s^2 \mu}, \tag{1}$$

where $b$ is the formation factor (usually between 10 and 20), $A_s$ the specific contact surface between grains and water [$m^2/m^3$], $\mu$ the water viscosity fixed at 0.0027 [kg/m s], $\rho$ the water density fixed at 999.7 [$kg/m^3$] (for freshwater at 10°C) and $g$ the gravity acceleration, 9.81 [$m^2/s$]. This relationship has been investigated for the USM formation and available statistical data are shown in Table 1.

| **Facies** | **Proportion** | $bA_s$ | **Mean** $\log_{10}K$ | $\sigma\log_{10}K$ | **Mean** $\phi$ | **Variance** $\phi$ |
|---|---|---|---|---|---|---|
| RG | 0.25 | 44000 | -5.95 | 1.46 | 0.209 | 0.003 |
| DFR | 0.36 | 166100 | -7.68 | 1.70 | 0.154 | 0.005 |
| UW | 0.10 | 700000 | -9.10 | 1.21 | 0.135 | 0.003 |
| UPS | 0.29 | 1200000 | -9.56 | 0.38 | 0.112 | 0.003 |

**Table 1** Summary of the parameters used to build the joint pdf displayed in figure 2 for the Kölliken site (Mariethoz, et al., 2009)

The second physical law is Archie's law (2), relating the electrical resistivity of the fluid saturated rock $R_t$ [Ωm] to the porosity (Archie, 1942):

$$R_t = R_w \phi^{-m}, \tag{2}$$

where $R_w$ is the fluid electric resistivity that is estimated to be around 5 [Ωm] in this formation (Hug, 2005) and $m$ is traditionally defined as a constant usually varying between 1.5 and 2.5, depending on the geometry of the pores.

The joint law is estimated by drawing samples of $\phi$ in the distributions presented in Table 1. For each $\phi$ sample, the couple $[K(\phi), R_t(\phi)]$ is computed. As $m$ is unknown, it is randomly sampled in a uniform distribution whose bounds are 1.5 and 2.5. A large number of samples are drawn until their number is sufficient to obtain a stable joint distribution by kernel smoothing (Epanechnikov, 1969). Figure 2a shows the resulting joint pdf, using 1,000,000 porosity samples. The specific characteristics of the different facies listed in table 1 are visible in the different modes of the marginal pdf of hydraulic conductivity (Figure 2b). The marginal distribution of $R_t$ (Figure 2c) corresponds to resistivity values measured in this part of the Molasse formation (Jin*, et al.*, 1995; Mächler, 1994).

This example shows that even if we assume very simple physical laws and statistical distributions for the porosity and basic parameters of these laws, we obtain a statistical relation between the two variables of interest which is complex and characterized by certain couples of values of log resistivity and hydraulic conductivity that are impossible while others are acceptable or even highly probable. For a given value of the resistivity such as 2, the bivariate distribution shows that a log hydraulic conductivity of -9 or -5.5 is acceptable but not a value of -7.5. It shows that it is possible to build the joint pdf with the proposed technique, but more important it shows also that multi-valued relations can occur from simple physical relations and from a mixture of different rock types at a small scale. We would like also to emphasize that when considering only a single rock type, the joint relation which is obtained using that approach is often extremely different from a multiGaussian distribution. There are, for example, very sharp boundaries between zones where some values are impossible and zones where the values have a certain probability of occurrence. We can expect that this type of relations is certainly very difficult to identify from a small number of samples that include measurement errors and this is why most often a multiGaussian model is used while it is most probably inadequate.

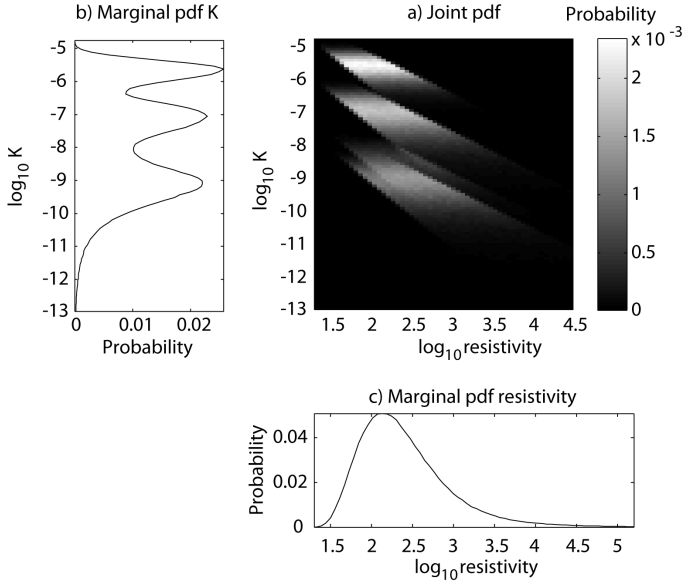



**Figure 2** Inference of the joint distribution a) The joint pdf resulting of porosity sampling. b) The marginal distribution of $K$. c) The marginal distribution of $R_t$.

## 3. Simulation by probability aggregation

### 3.1. Outline of the method

Before discussing the details of the method, let us outline its main characteristics. First, it is a sequential simulation algorithm. The nodes are successively visited in a random order. For each successive node **x**, two local conditional cumulative distribution functions (ccdf) are estimated. One, $F_1(\mathbf{x},z)$, is the distribution function of the main attribute $Z(\mathbf{x})$ conditional to the neighboring data (previously simulated nodes and conditioning data) and the spatial correlation model:

$$F_1(\mathbf{x};z) = \text{Prob}\{Z(\mathbf{x}) \le z \mid z(\mathbf{x}_1),...,z(\mathbf{x}_N)\}. \quad (3)$$

This conditional ccdf can be estimated using any suitable geostatistical method. In our examples, multiGaussian kriging is used.

At the same location **x**, the distribution function, $F_2(\mathbf{x};z|s)$, of $Z(\mathbf{x})$ conditional to the co-located attribute $s(\mathbf{x})$ can be obtained in a straightforward manner because $s(\mathbf{x})$ is exhaustively known. One just needs to extract it from the 2D bivariate joint probability density function *f(z,s)*:

$$F_2(\mathbf{x};z|s) = \text{Prob}\{Z(\mathbf{x}) \le z | S(\mathbf{x}) = s\} = \frac{\int_{\varsigma=-\infty}^{\varsigma=z} f(\varsigma, s(x))\, d\varsigma}{\int_{\varsigma=-\infty}^{\varsigma=+\infty} f(\varsigma, s(x))\, d\varsigma}. \quad (4)$$

$F_1(\mathbf{x};z)$ and $F_2(\mathbf{x};z|s)$ provide two distinct pieces of information on the value that should be finally assigned to $Z(\mathbf{x})$. The issue is therefore to combine (aggregate) $F_1(\mathbf{x};z)$ and $F_2(\mathbf{x};z|s)$ into a single ccdf $F(\mathbf{x};z|s)$ which would be an approximation of:

$$\text{Prob}\{Z(\mathbf{x}) \le z | z(\mathbf{x}_1),...,z(\mathbf{x}_N), S(\mathbf{x}) = s\}. \quad (5)$$

Once this ccdf is available for the location **x**, a value is drawn from it and assigned to $Z(\mathbf{x})$. As usual in sequential simulation, $Z(\mathbf{x})$ is thereafter treated as conditional data for simulating the remaining unknown values of *Z*.

## 3.2. Estimating $F_1(\mathbf{x};z)$ using multiGaussian kriging

Estimating a local ccdf $F_1(\mathbf{x};z)$ given a set of conditioning data $z(\mathbf{x}_i)$, $i=[1...N]$ can be achieved by a variety of geostatistical techniques. MultiGaussian kriging (Emery, 2005; Goovaerts, *et al.*, 2005; Verly, 1993) is probably the most widely used and well suited in the case of high entropy phenomena (as opposed to more structured, low entropy phenomena, that sould be described with methods such as Multiple-Point statistics). The main advantage of MultiGaussian kriging is that it allows estimating for each location the complete pdf of the variable of interest even if its univariate distribution is not Gaussian. The approach consists of:

- Performing a normal score transform of the conditioning data: $y(\mathbf{x}) = G[z(\mathbf{x})] \sim N\begin{cases}0\\1\end{cases}$.
- Assuming that this transformation is sufficient to ensure multi-Gaussianity.

- Estimating at node **x**, by simple kriging, the mean and variance of the local conditioning Gaussian distribution $m$, $\sigma^2$.
- Back-transformation of the entire local Gaussian distribution in order to obtain a distribution function that is not necessarily Gaussian. This is done numerically by applying the back-transformation $G^{-1}[y(\mathbf{x})]$ on equally spaced quantiles of the distribution found by simple kriging. The non-Gaussian local distribution is then reconstructed from its quantiles.

## 3.3. Probability Aggregation

The problem of aggregating $F_1(\mathbf{x};z)$ and $F_2(\mathbf{x};z|s)$ is not trivial. Bayesian updating (e.g. Woodbury and Ulrych, 2000) could be used in this context, but it assumes conditional independence. Therefore, it would require some knowledge of the relationship between the secondary variable at the current location $S(\mathbf{x})$ and the primary variable at all known locations, including all previously simulated nodes. According to Journel (2002), conditional independence is generally too strong an assumption in the context of sequential simulation. Indeed, the ccdfs defined in equations (3) and (4) are not conditionally independent because $F_1(\mathbf{x};z)$ is based on previously simulated nodes that already integrated information on the joint distribution $f(z,s)$.

Management science provides methods for aggregating expert's opinions while dealing with data interaction and without assuming conditional independence. Bordley (1982) proposed a formula for computing an aggregated probability density function $f^a(z)$ from $n$ individual pdfs $f^k(z)$, $k=[1\ldots n]$ (representing the various expert opinions) that satisfies the basic probability properties and a series of axioms, including the weak likelihood ratio axiom:

$$f^a(z) = \frac{1}{\eta} f^0(z) \prod_{k=1}^{n} \left[ \frac{f^k(z)}{f^0(z)} \right]^{w_k}, \tag{6}$$

where $\eta$ is a normalizing factor. Each probability density function has a weight $w_k \in \Re$, which can be seen as a way of quantifying redundancy or as a confidence factor. $f^0$ is the prior density function which is, in our case, the marginal pdf $f(z) = \int_{s=-\infty}^{s=-\infty} f(z,s)\,ds$ (i.e. the only information available on $z$ when $s$ is unknown). By

construction, Bordley's formula has a certain number of important mathematical properties (see Clemen and Winkler, 2007 for a discussion), but nevertheless it is only an approximation allowing to combine several probabilities of the same event to occur (estimated from different sources of information) when the complete probability model including all the sources of information, and all the data interaction, is lacking. Equation (6) is closely related to the *tau* and *nu* models (Journel, 2002; Krishnan, 2005; Polyakova and Journel, 2007). In the same spirit, Tarantola (2005) defines the conjunction of probability densities as the simplest way to aggregate probabilities in that manner:

$$f(\mathbf{x}; z \mid s) = f_1(\mathbf{x}; z) \wedge f_2(\mathbf{x}; z \mid s) = \frac{1}{\eta} \frac{f_1(\mathbf{x}; z)\ f_2(\mathbf{x}; z \mid s)}{f(z)}. \tag{7}$$

Equation (7) is a particular case of equation (6) with two probabilities being aggregated with identical weights equal to 1. Nevertheless, such restrictions are not necessary. Using (6), it is possible to apply the method on any number of attributes and to adjust $w_k$ in order to assign the relative weights to an expert-provided bivariate distribution model or to account for a spatially variable model of uncertainty by setting different weights values at different locations. Setting a weight to 0 at a certain location would result in the corresponding source of information having no influence.

Our purpose is to aggregate $f_1(\mathbf{x}; z)$ and $f_2(\mathbf{x}; z \mid s)$ and we exposed various ways of achieving it. We do not want to favor a specific method among the ones mentioned above. In our examples, we illustrate the method using equations (6) and (7), but there are no restrictions regarding other techniques.

## 3.4. Step-by-step algorithm

The proposed algorithm is implemented as follows:

1. Define the marginal cdf and spatial correlation model of $Z(\mathbf{x})$ as well as the bivariate model $f(z,s)$.
2. Each conditioning data is assigned to the closest grid node in the simulation grid (SG).
3. Define a path through the remaining nodes of the SG. The path is a vector containing all the indices of the grid nodes that will be simulated

sequentially. Any type of path is suitable, for example random or unilateral (Daly, 2004; Pickard, 1980).

4. For each successive location **x** in the path:
    a. Infer the local pdf $f_1(\mathbf{x};z)$ from the known conditioning data in the neighborhood of **x**. Any appropriate geostatistical method can be used for this purpose. In the following examples, we use multiGaussian kriging (see above for details).
    b. Extract $f_2(\mathbf{x};z|s)$ from the bivariate model (that can be spatially-dependent or not).
    c. Estimate $f(\mathbf{x};z|s)$ by probability aggregation using Bordley's equation (6) or using the conjunction of probability (7) (i.e using weigths equal to one). In the examples below we will first start with the conjunction of probability and later test the effect of varying the weights. Note that the *tau* or *nu* models could also be used here to aggregate $f_1(\mathbf{x};z)$ and $f_2(\mathbf{x};z|s)$ in order to obtain $f(\mathbf{x};z|s)$.
    d. Randomly draw a sample $z'(\mathbf{x})$ from $f(\mathbf{x};z|s)$, assign it to the location **x** in the grid and add it to the conditioning data set.

## 4. Synthetic example

The proposed algorithm has was on a synthetic example. For this example, a bivariate density function is known and a synthetic reference field for the primary attribute is created by unconditional simulation. The secondary attribute is constructed from the (fully known) primary attribute by drawing for each node **x** a value from

$$F(\mathbf{x};s\,|\,z) = \text{Prob}\{S(\mathbf{x}) \le s\,|\,Z(\mathbf{x}) = z\}\,. \tag{8}$$

Then, the reference is sampled at 50 random locations. This is used as input conditioning data for the algorithm. The simulation grid size is 50x50 cells, and 100 realizations are generated. In order to evaluate the performance of the method, the simulations are compared to the reference which is known exhaustively. The comparison criteria are the reproduction of the histogram and variogram, the errors between the simulated values and the known reality, and the visual aspect of the simulations.

Figure 3 illustrates the method. The bivariate distribution function is inspired from the Oman case described in the introduction. The primary attribute (Figure 3a) has a bimodal distribution (Figure 3g). Locations of randomly sampled data are marked by crosses (Figure 3a). The secondary attribute is noisy (Figure 3b) and is related to the primary attribute by a crescent-shaped joint pdf (Figure 3c). Despite the noise, the secondary attribute still contains enough information to guide the simulations (Figure 3d and 3e), where features of the reference are present at locations where no data are available (for example the dark channel that runs through the field from left to right).

The joint distribution (Figure 3f), the reference histogram (Figure 3g) and variogram (Figure 3h) are well reproduced (the solid red line represents the reference, dots represent the simulations and blue circles the sampled data). There is no systematic bias, as shown by the histogram of errors of the simulated attribute that is centered on 0 (Figure 3i).

By construction, the relation between the secondary and primary variable is non-injective. As expected the proposed method allows to generate an ensemble of simulations of the primary field that respect this very particular relation as shown by the reproduction of the the joint pdf (Figure 3f). Among the simulations, at locations where one finds low values of the secondary variable, the primary variable is either extremely low or extremely high.

In order to test whether the method could also be applied in more standard cases where a linear and multiGaussian relation holds, several tests were made with different correlation coefficients and variograms for the primary attribute. The accuracy of the method was compared to standard cosimulation algorithm. These tests are not shown in this paper for the sake of brevity, however they have all shown that the method performed as well as a traditional collocated cosimulation method. However, there is no advantage of using our method in those cases. When the multiGaussian assumption holds, it is wiser to use a full cosimulation or cokriging based method because it accounts for the value of the secondary variable in the entire neighborhood, and not only at the location to simulate. Moreover, it takes advantage of a fully consistent model, whereas certain parameters of probability aggregation are subject to calibration (e.g. the weights of equation 6).

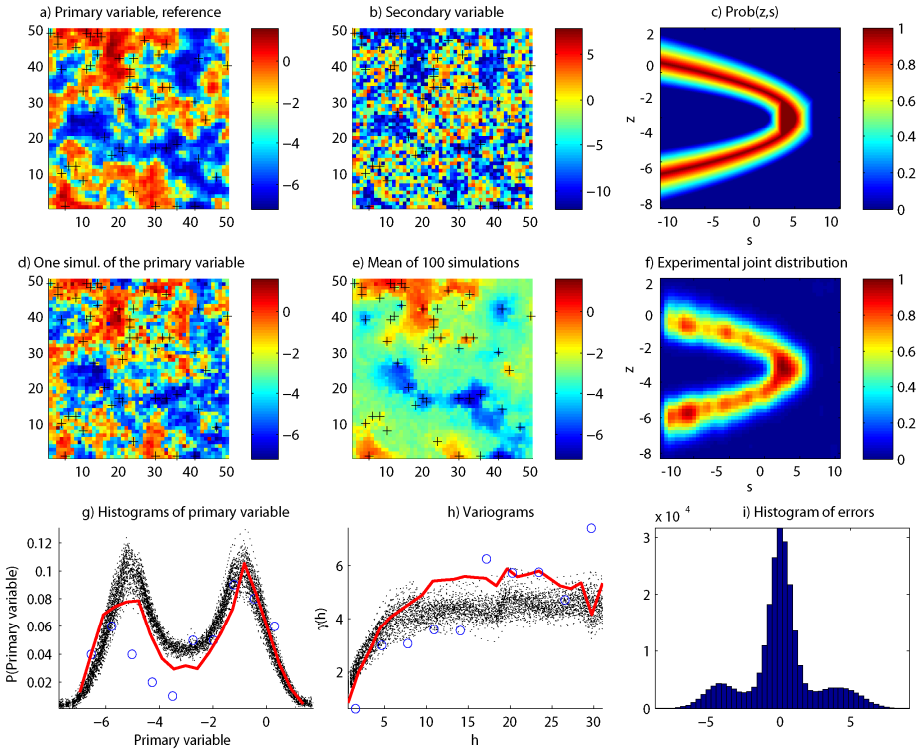



**Figure 3** Synthetic example using a multi-valued relation modeled by a custom joint pdf. The primary attribute has a spherical variogram model (sill=5.3, range=12, adjusted on the 50 sample data).

## 5. Realistic example

In order to test further the method, a real data set was used. It is based on two Landsat 7 satellite sensor images corresponding to the same area, but taken at two different wavelengths (Figure 4a and 4b). One image is considered to be the primary attribute while the other is the secondary attribute. We decided to use such images first because satellite sensor images are often used as secondary variables in hydrology, for example for the estimation of soil moisture (Makkeasorn, *et al.*, 2006; Zribi, *et al.*, 2005) or for adding accuracy to the mapping of ground based measurements of precipitation (Haberlandt, 2007). But more importantly two satellite images constitute a unique data set of two exhaustively known variables allowing to test the accuracy of the method with real data and not on a synthetic case. Indeed, exhaustive data sets are seldom available for other variables that may

be of greater interest in groundwater hydrology (exhaustive hydraulic conductivity map and exhaustive geophysical survey for example). Of course, in practice we are generally not interested in estimating a satellite image at a given wavelength knowing the image at another wavelength and at a few discrete locations. The aim here is only to test the method on real data. Because different processes affect the absorption of light at different wavelengths, each image highlights different features of the land and the joint relationship is thus complex (Figure 4c). In summary, this data set constitutes an analogue to the real problems encountered in practice, but an analogue whose primary variable is exhaustively known and therefore an analogue allowing to evaluate the accuracy of the method.

As for the synthetic examples, multiGaussian kriging is used for estimating $F_1(\mathbf{x};z)$. The first image is used as a reference, sampled at 100 random locations, while the second is the auxiliary attribute. The size of the simulation grid is 181x201. Figures 4d to 4h present the results of the 100 simulations in the same fashion as the synthetic examples. Figure 4i shows the standard deviation of the stack of 100 simulations. The joint distribution used for the cosimulation (Figure 4c) is very well reproduced in the simulations (Figure 4f). The root mean square error (RMSE) of the simulated values compared to the reference field is 24.70.

To compare the method against existing and well established cosimulation methods, we used the traditional collocated cosimulation technique (Almeida and Journel, 1994) to generate 100 simulations of the primary variable with the same input data: the same exhaustive secondary variable, the same co-located data points, adjusted cross-variograms, plus the assumption of a linear correlation between the Gaussian transformed variables. The method was applied with care and all necessary Gaussian direct and back transformations were performed. The results are showed in figures 4j to 4o, in the same manner as the previous figures, and the validation criteria are the same.

The traditional cosimulation method provides good variogram and histogram reproduction, even if the reproduction is less than what was obtained with probability aggregation. The RMSE compared to the reference field is 29.46. Nevertheless, the reproduction of the bivariate joint probability density function is grossly inaccurate (Figure 4l). This is due to the violation of the assumption of linear correlation between both variables. The Gaussian transformations result in a relationship that is not-linear, but still very far from the true relationship (Figure 4c).

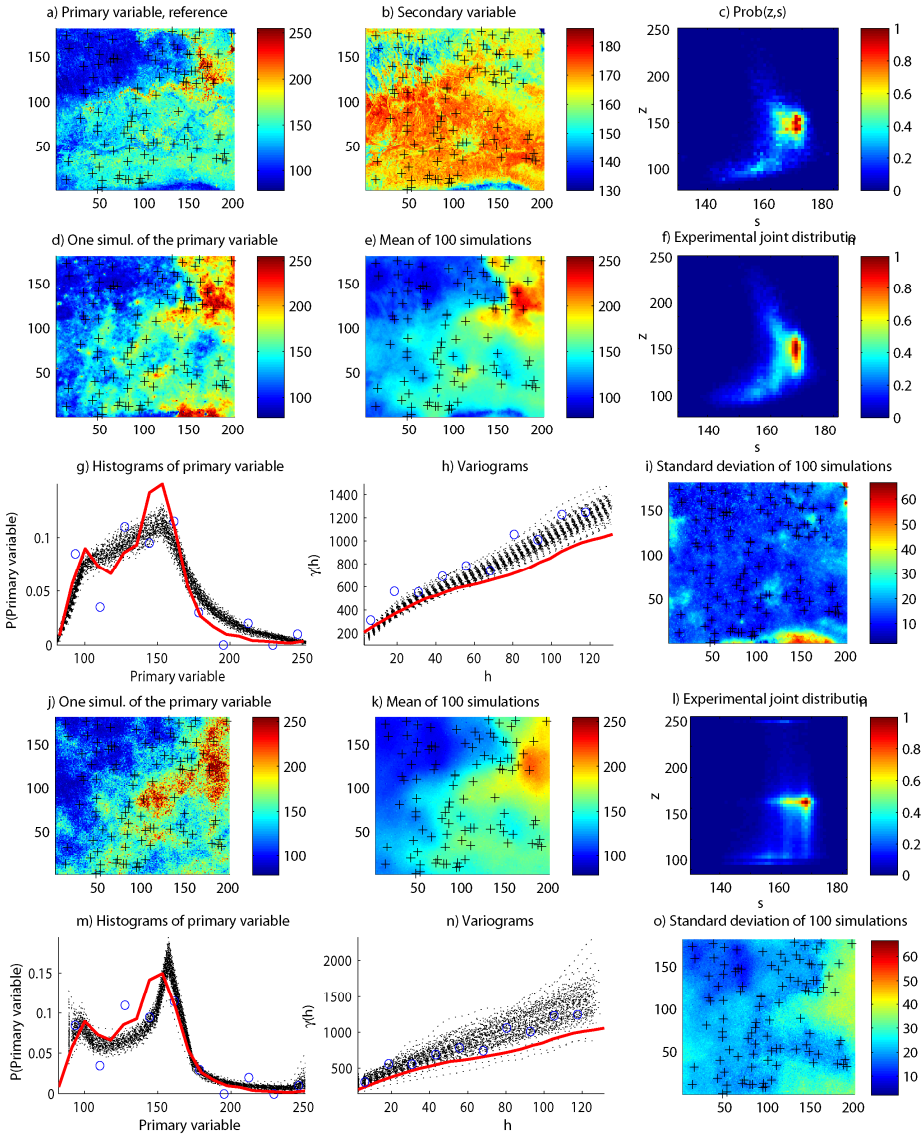



**Figure 4** Real case example. a) to c) Input data. The primary attribute has an exponential variogram model (sill=1110, range=90, adjusted on the 100 sample data). d) to i) Results and validation criteria using the probability aggregation method. j) to o) Results and validation criteria using collocated co-simulation, with the same input data.

In terms of reproduction of the reference image, probability aggregation is able to reproduce detailed features (visible on the mean of the simulations Figure 4e) that are too specific to be inferred using only the primary variable data and its variogram. Only the secondary variable contains such detailed local information, but as the relationship is non-linear, the traditional cosimulation approach is unable to provide a similar level of detail (compare Figures 4d and 4j to the reference, Figure 4a).

The comparison of the standard deviation maps (Figures. 4i and 4o) show how much information probability aggregation is able to extract from the secondary variable. Standard deviation map of traditional cosimulation (Figure 4o) is mainly related to the distance to the data points (high standard deviation when no data are present). Standard deviation maps issued from probability aggregation also show high uncertainty at locations where the secondary information carries a low information content (e.g. the lower part of the image where the secondary variable value can correspond to a wide range of values for the primary variable). At locations where the secondary variable is very informative, standard deviation is low, even in the absence of conditioning data.

## 6. Adjusting the weights

So far, both weights $w_1$ and $w_2$ have been kept identical and equal to 1. This is justified when the confidence related to each source of information as well as the data interaction are unknown (Polyakova and Journel, 2007). Nevertheless, using equation (6) allows setting these parameters to modify the influence of each source of information. In the example described above, these parameters can be adjusted in order to fit the simulation on the reference field.

A sensitivity analysis on $w_1$ and $w_2$ was carried out by testing all pairs $[w_1,w_2]$, with each parameter varying from 0 to 3 with a step of 0.2, thus resulting in 256 possible pairs. A stack of 10 realizations of the primary attribute was generated for each pair of weights. The mean error of each stack compared to the reference was calculated. The results of this sensitivity analysis are shown graphically in figure 5. For each stack, this error was evaluated using 4 criteria, consisting in the mean sum of squared differences between the reference and the simulations of the stack for the histogram (Figure 5a), the variogram (Figure 5b), the joint pdf (Figure 5c) and the values of the primary attribute (Figure 5d) at each grid node.

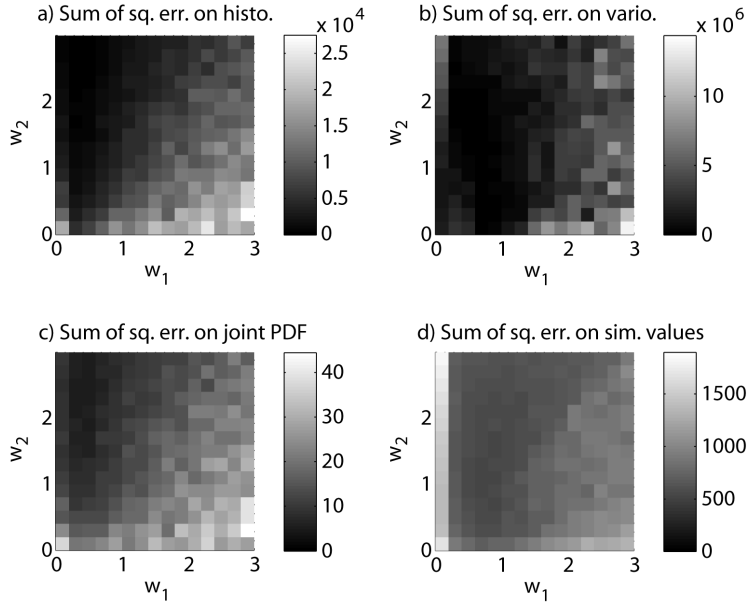



**Figure 5** Sensitivity analysis of the value of the weights $w_1$ and $w_2$.

Good fits are obtained by setting $w_2$ to a high value (about 2) and $w_1$ to a low value (about 0.5). This emphasizes the high local information content of the secondary attribute when the joint law is accurately estimated. But using only $F_2(\mathbf{x};z|s)$ (setting $w_1$ to 0) generates a result that depends only on the joint law, and which might be biased if this law is inaccurate. The information represented by $F_1(\mathbf{x};z)$, although partially redundant with $F_2(\mathbf{x};z|s)$, is also capital because it ensures a spatial consistency in the simulated field. Indeed, setting $w_1$ to 0 dramatically decreases the quality of the simulation. Moreover, $F_1(\mathbf{x};z)$ contains both local and structural information, whereas $F_2(\mathbf{x};z|s)$ contains local information only.

This sensitivity analysis shows that the optimal weights depend on various factors that are difficult to foresee in practice if the true primary variable field is unknown. As a workaround, we propose to determine the goodness of a pair $[w_1,w_2]$ using cross-validation (e.g. Dubrule, 1983). The true and estimated values can be compared in several ways: constructing the bivariate plot of true versus estimated values, building the histogram of their differences or, simply, by mapping the differences. All these representations give us ways to assess the adequacy of the model. Two caveats, though: cross-validation cannot tell that the model is right; it

will only highlight its deficiencies. Secondly, in all rigor, the data to be used in the cross-validation stage should not be used for building the model.

In practice, to determine the goodness of a pair $[w_1,w_2]$ using cross-validation, we propose to proceed as follows:

1. At the location $\mathbf{x}_i$ of each data point:
    a. Infer $f_1(\mathbf{x}_i;z)$ as previously, but without accounting for the fact that $z(\mathbf{x}_i)$ is actually known.
    b. Extract $f_2(\mathbf{x}_i;z|s)$ from the bivariate model.
    c. Estimate $f(\mathbf{x}_i;z|s)$ by probability aggregation using a given pair of weights $[w_1,w_2]$.
    d. Estimate $P^* = f(\mathbf{x}_i;z(\mathbf{x}_i)|s)$, the probability associated to the true value $z(\mathbf{x}_i)$.

Compute the mean probability of all data values $P_m^* = \frac{1}{N}\sum_i P^*(z(\mathbf{x}_i))$. In itself, this mean probability is not very informative, but it allows comparing various pairs of weights and determining which one yields better results (i.e. where the true values are the most probable).

$P_m^*$ has been computed the same pairs of weights $[w_1,w_2]$ as the sensitivity analysis. The results are displayed in Figure 6. The optimal weights found with cross-validation are very similar to those found previously, but this time they were computed using the available data set only. This is very important for practical applications, where the reference field is never available.

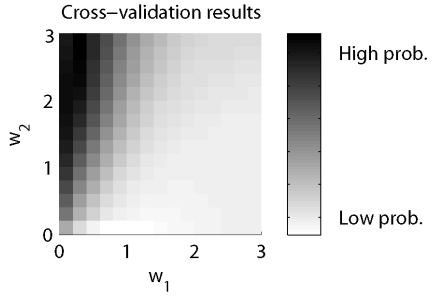



**Figure 6** Optimal weights found by cross-validation. The color scale represents the average probability of all measured data given by the model when a certain pair of weights $[w_1,w_2]$ is used.

100 additional simulations were generated using the weights obtained by cross validation, and the same input data as previously. Results are presented in Figure 7. The reference field is very well reproduced, with highly accurate histograms and variograms fit, good reproduction of the bivariate distribution and a RMSE of 25.37. The features of the reference field are well reproduced. More interesting is the standard deviation map, where all features specific to the secondary variable are highlighted. This is because more weight has been put on $w_2$, thus generating high variability when $F_2$ is not well determined.

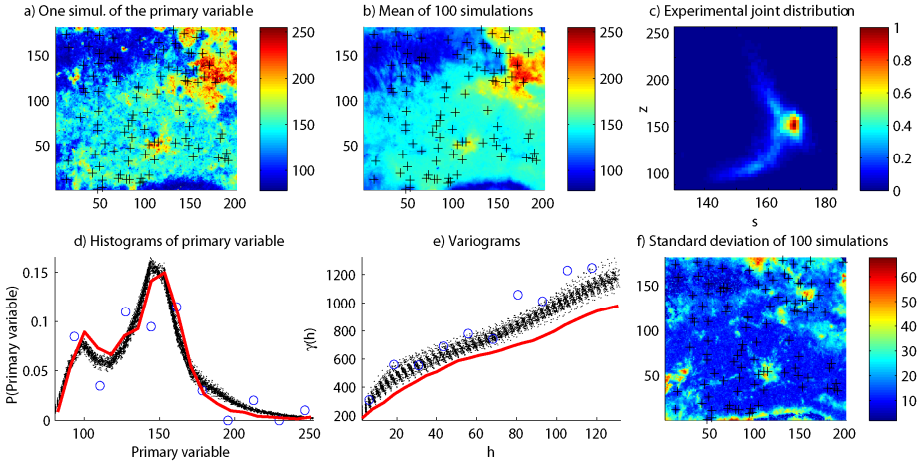



**Figure 7** Real case example using probability aggregation with optimal weights ($w_1$=0.5, $w_2$=2). The input data are the same as previously.

## 7. Discussion and conclusion

Field observations made in a coastal aquifer in Oman (Alcolea, *et al.*, 2009), statistical calculations using simple physical laws (Figure 2), or the relation between the magnitude of the signals observed at different wavelength on satellite images (Figure 4c) suggest that non multiGaussian and possibly non injective relations need to be accounted for in a significant number of situations in surface and groundwater hydrology when interpolating certain primary variables using an exhaustive map of secondary information. To account for this type of relations, the simplest and most general approach is to model it using a joint probability distribution. In the examples that were used to illustrate and test the proposed methodology, we used a non-parametric joint distribution that offers a high degree of flexibility. However, one

can also use an analytical expression for the joint distribution when such a model is available without any change in the algorithm.

The proposed geostatistical simulation algorithm is an extension of the collocated cosimulation techniques. Its originality is that it is not based on an analytical and explicit model of the relationship between the various sources of information. Instead, it uses the joint probability density distribution to express at any location the conditional distribution of the primary attribute knowing the secondary attribute and a model of spatial continuity for the primary attribute. These two models are assembled by a weighted probability aggregation technique.

The main advantage of this method is that it allows to provide more accurate maps of the primary attribute than standard techniques when an exhaustive map of secondary information is available and when the relation between the two variables is better described by a joint probability distribution. Many hydrological applications could then benefit from that method. Another advantage is that it can be extended easily to multiple sources of information. A first possible extension is to use not only two local pdfs but as many local pdfs as needed. Adjusting the weights $w_k$ is then a powerful way of parametrizing the method to aggregate secondary attributes having different information content.

The drawback of this flexibility is that finding the appropriate weights can be difficult. Keeping the weights equal to 1 does not assume conditional independence, but instead assumes the absence of data interaction, which has fewer consequences. However, a proper way of adjusting them is desired. This is made difficult because the weights describe at the same time both concepts of confidence and redundancy. Polyakova and Journel (2007) suggest two methods to determine them: the first one is to calibrate the weights using available data, and the second one is to determine them using proxy cases. In our example, the first method gave similar weights to the ones found with a complete sensitivity analysis including a full knowledge of the true reference field. This tends to show that weights can be calibrated satisfactorily if enough hard data are present.

An issue that was not considered in this work is that the relationship between the primary and the secondary variables can vary spatially (Day-Lewis and Lane, 2004) or as a function of a third or fourth auxiliary variables. A strength of our approach is that this additional information, when it is known, can be modeled using not only a bivariate joint probability distribution but a $n$-dimensional probability cube. At each location, one could compute the conditional probability density distribution of the primary variable knowing all the secondary variables known at that location. The rest of the method would remain unchanged. In doing so the method would

accommodate spatially-dependent statistical relationship between variables and any number of auxiliary attributes.

One main limitation of the method is that it does not include an explicit model for joint spatial cross-correlations between primary and secondary attributes. Only the spatial correlation of the primary attribute is modeled. We believe that this limitation is compensated by the flexibility of adjusting individual weights for an unlimited number of secondary attributes, and by the simplicity of the algorithm.

The method has been tested using a multiGaussian model for the spatial continuity of the primary variable, but it can directly be extended to any sequential, pixel-based simulation technique that uses local conditional pdfs. Moreover, it is not limited to continuous attributes (for example, it can be used in the framework of multiple points statistics). Therefore, its straightforward implementation makes it interesting to append probability aggregation on existing simulation codes.

Finally, this paper has shown that the concept of probability conjunction or aggregation, originating from management science, is a precious tool for integrating information stemming from diverse sources in problems related to the characterization of hydrological processes.

## 8. References

Ahmed, S., and Demarsily, G. (1987), *Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity*, Water Resour. Res., 23, 9, 1717-1737.

Alcolea, A., Renard, P., Mariethoz, G., and Bretone, F. (2009), *Reducing the impact of a desalination plant using stochastic modeling and optimization techniques*, Journal of Hydrology, 365, 3-4, 275-288.

Almeida, A., and Journel, A. (1994), *Joint Simulation of Multiple Variables with a Markov-Type Coregionalization Model*, Mathematical Geology, 26, 565-588.

Archie, G. (1942), *The Electrical Resistivity Log as an Aid in Determining Some Reservoir Characteristics* J. Pet. Technol., 5, 1-8.

Bashore, W., U.Araktingi, Levy, M., and Schweller, U. (1994), *Importance of a Geological framework for Reservoir modelling and subsequent Fluid-Flow Predictions*, in *AAPG Computer application in geology*, pp. 159-175

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, John Wiley & Sons, Inc.

Bordley, R. E. (1982), *A multiplicative formula for aggregating probability assessments*, Management Science, 28, 10, 1137-1148.

Brunner, P., Bauer, P., Eugster, M., and Kinzelbach, W. (2004), *Using remote sensing to regionalize local precipitation recharge rates obtained from the Chloride Method*, Journal of hydrology, 294, 4, 241-250.

Caers, J. (2001), *Automatic Histogram and Variogram Reproduction in Simulated Annealing Simulation*, Mathematical Geology, 33, 2, 167-190.

Cassiani, G., Böhm, G., Vesnaver, A., and Nicolich, R. (1998), *A geostatistical framework for incorporating seismic tomography auxiliary data into hydraulic conductivity estimation*, Journal of hydrology, 206, 1-2, 58-74

Chilès, J.-P., and Delfiner, P. (1999), *Geostatistics - Modeling Spatial Uncertainty*, John Wiley & Sons, Inc.

Clemen, R. T., and Winkler, R. L. (2007), *Aggregating probability distributions*, in *Advances in decision analysis: from foundations to applications*, pp. 154-176, Cambirdge University Press.

Creutin, J. D., Delrieu, G., and Lebel, T. (1988), *Rain measurement by raingage-radar combination: a geostatistical approach*, Journal of Atmospheric and Oceanic Technology|Journal of Atmospheric and Oceanic Technology, 5, 1, 102-115.

Dafflon, B., Irving, J., and Holliger, K. (2008), *Simulated-annealing-based conditional simulation for the local-scale characterization of heterogeneous aquifers* Journal of Applied Geophysics.

Dagan, G. (1989), *Flow and transport in porous formations*, Springer.

Daly, C. (2004), *Higher order models using entropy, Markov random fields and sequential simulation*, paper presented at Geostatistics Banff 2004.

Day-Lewis, F., and Lane, W. (2004), *Assessing the resolution-dependent utility of tomograms for geostatistics*, Geophys. Res. Lett., 31, L07503.

Delhomme, J.-P., Boucher, M., Meunier, G., and Jenson, J. (1981), *Apport de la géostatistique à la description des stockages de gaz en aquifère*, *Revue de l'Institut Français du Pétrole*, 36, 3, 209-327.

Deutsch, C. (1992), *Annealing techniques applied to reservoir modeling and the integration of geological and engineering (well test) data*, Ph.D Thesis, Stanford University.

Dubrule, O. (1983), *Cross-validation of kriging in a unique neighborhood*, Mathematical Geology, 15, 7, 687-699.

El Idrisy, E., and De Smedt, F. (2007), *A comparative study of hydraulic conductivity estimations using geostatistics*, Hydrogeology Journal, 15, 3, 459-470.

Emery, X. (2005), *Simple and Ordinary Multigaussian Kriging for Estimating recovearble Reserves*, Mathematical Geology, 37, 3, 295-319.

Epanechnikov, V. A. (1969), *Nonparametric estimation of a multidimensional probability density*, Theoretical Probability Applications, 14, 153-158.

Gomez-Hernandez, J., and Journel, A. (1993), *Joint Simulation of MultiGaussian Random Variables*, in *GeostatisticsTroia '92, vol.1*, pp. 85-94, Kluwer Academic Press.

Goovaerts, P. (2000), *Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall*, Journal of hydrology, 228, 1-2, 113-129.

Goovaerts, P., AvRuskin, G., Meliker, J., Slotnick, M., Jacquez, G., and J., N. (2005), *Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan*, Water Resour. Res., 41, 7, W07013.

Haberlandt, U. (2007), *Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event*, Journal of hydrology, 332, 1-2, 144-157.

Hu, L., and Chugunova, T. (2008), *Multiple-Point Geostatistics for Modeling Subsurface Heterogeneity: a Comprehensive Review*, Water Resour. Res., 44, W11413.

Hug, R. (2005), *Hydrogeologische Untersuchungen im Abstrom der Sondermülldeponie Kölliken*, Bull. angew. Geol., 10, 1, 65.

Izenman, A. (1991), *Recent Developments in Nonparametric Density Estimation*, J. Am. Stat. Assoc., 86, 413, 205-224.

Jin, J., Aigner, T., Luterbacher, H. P., Bachmann, G. H., and MLiller, M. (1995), *Sequence stratigraphy and depositional history in the south-eastern German Molasse Basin*, Marine and Petroleum Geology, 12, 8, 929-940.

Journel, A., and Alabert, F. (1990), *New Method for Reservoir Mapping*, Journal of Petroleum Technology, SPE paper 20781.

Journel, A., and Isaaks, E. (1984), *Conditional indicator simulation: Application to a Saskatchewan deposit*, Mathematical Geology, 16, 7, 685–718.

Journel, A. G. (2002), *Combining Knowledge From Diverse Sources: An Alternative to Traditional Data Independence Hypotheses*, Mathematical Geology, 34, 5, 573-596.

Kolbjørnsen, O., and Abrahamsen, P. (2004), *Theory of the cloud transform for applications*, in *Geostatistics Banff 2004*, pp. 45-54, Kluwer Academic Publisher.

Krishnan, S. (2005), *Combining diverse and partially redundant information in the earth sciences*, Ph.D Thesis, Stanford University.

Leuangthong, O., and Deutsch, C. (2003), *Stepwise Conditional Transformation for Simulation of Multiple Variables*, Mathematical Geology, 35, 2, 155-173.

Mächler, E. (1994), *Kombination verschiedener geophysikalischer Methoden (Refraktionsseismik und Widerstandsgeolektrik) zur Aufsuchung einer Molasserinne in Kölliken (AG)*, Diploma Thesis, ETH Zürich.

Makkeasorn, A., Chang, N., Beaman, M., Wyatt, C., and Slater, C. (2006), *Soil moisture estimation in a semiarid watershed using RADARSAT-1 satellite imagery and genetic programming*, Water Resour. Res., 42, W09401.

Mariethoz, G., Renard, P., Cornaton, F., and Jaquet, O. (2009), *Truncated plurigaussian simulations to characterize aquifer heterogeneity*, Ground Water, 47, 1.

Matheron, G. (1971), *The theory of regionalized variables and its application.*, Ecoles des Mines de Paris.

Moysey, S., and Knight, R. (2004), *Modeling the field-scale relationship between dielectric constant and water content in heterogeneous systems*, Water Resour. Res., 40, W03510.

Moysey, S., Singha, K., and Knight, R. (2005), *A framework for inferring field-scale rock physics relationships through numerical simulation*, Geophys. Res. Lett., 32, L08304.

Mukerji, T., Jorstad, A., Avseth, P., Mavko, G., and Granli, J. (2001), *Statistical rock physics: Combining rock physics, information theory, and geostatistics to reduce uncertainty in seismic reservoir characterization*, The Leading Edge, 20, 3, 313-319.

Ortiz, J. M., and Deutsch, C. V. (2004), *Indicator Simulation Accounting for Multiple-Point Statistics*, Mathematical Geology, 36, 5, 545-565.

Pickard, D. (1980), *Unilateral Markov fields*, Advances in Applied Probability, 12, 655-671.

Polyakova, E., and Journel, A. (2007), *The Nu Expression for Probabilistic Data Integration* Mathematical Geology, 39, 8, 715-733.

Slater, L., and Lesmes, D. (2002), *Electrical-hydraulic relationships observed for unconsolidated sediments*, Water Resour. Res., 38, 10, 1213.

Soares, A. (2001), *Direct sequential simulation and cosimulation*, Mathematical Geology, 33, 8, 911-926.

Soupios, P., Kouli, M., Vallianatos, F., Vafidis, A., and Stavroulakis, G. (2007), *Estimation of aquifer hydraulic parameters from surficial geophysical methods: A case study of Keritis Basin in Chania (Crete – Greece)*, Journal of Hydrology, 338, 1-2, 122-131.

Tarantola, A. (2005), *Inverse Problem Theory and Methods for Parameter estimation*, Society for Industrial and Applied Mathematics.

Verly, G. (1993), *Sequential Gaussian Cosimulation - a Simulation Method Integrating several Types of Information*, in *Geostatistics Troia '92, Vol. 1*, pp. 543-554, Kluwier Academic Publishers.

Wackernagel, H. (2003), *Multivariate Geostatistics: an introduction with applications* 3rd edition ed., Springer-Verlag.

Winkel, L., Berg, M., Amini, M., Hug, S. J., and Johnson, C. A. (2008), *Predicting groundwater arsenic contamination in Southeast Asia from surface parameters*, Nat. Geosci., 1, 8, 536-542.

Woodbury, D., and Ulrych, T. (2000), *A full-Bayesian approach to the groundwater inverse problem for steady state flow*, Water Resour. Res., 36, 8, 2081-2093.

Xu, W., Tran, T., Srivastava, M., and Journel, A. G. (1992), *Integrating seismic data in reservoir in reservoir modeling: The collocated cokriging alternative. SPE Paper 24742*, paper presented at 67th SPE Annual Technical Conference and Exhibition.

Zribi, M., Baghdadi, N., Holah, N., and Fafin, O. (2005), *New methodology for soil surface moisture estimation and its application to ENVISAT-ASAR multi-incidence data inversion* Remote Sensing of Environment, 96, 3-4, 485-496.

# Chapter 7

# Concluding comments

*"The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work."*

*John Von Neumann*

## 1. Personal views

Hydrogeologists, geologists, engineers and modelers need images of physical processes. Images are crucial tools for interpretation, representation and modeling of these processes. The term of image is defined here in a very broad sense as an exhaustive representation of a process deployed in space / time. Such images can be either 1D (e.g. time series of piezometric level measurements), 2D (e.g. maps of precipitation intensity), 3D (e.g. representation of geological structures), or have higher dimensionality when both space and time variability are represented. The case of multivariate images can also be considered as an extra dimensionality.

Stochastic simulation methods generate images that are susceptible to mimic real-life processes. This is accomplished by isolating the statistical properties associated to the outcome of natural processes, and then by using these statistics as constraints for the generation of images. This is not without consequences. Representing a natural phenomenon by the statistics of its outcomes is a simplification. Natural processes are an elaborate combination of randomness and deterministic physical laws and, therefore, real variables fields present complex idiosyncrasies. Statistical language often describes this complexity with terms such as “non-stationarity”, “non-heteroscedasticity” or “non-Gaussianity”. In my opinion, the negative character of these terms tells how statistics fail to describe this complexity. Nevertheless, images are needed, no matter if they are not entirely satisfactory.

The most successful attempts at reproducing realistic complexity were accomplished by multiple-points and patterns simulations methods that allow increasing dramatically the order of the statistics by using non-parametric models. Nevertheless, several concerns are often raised against these methods, such as:

- The underlying random function of the generated fields is undefined.
- The variability between realizations is implementation-dependent (Journel and Zhang, 2006), and depends on algorithmic parameters such as the neighborhood size.
- It is practically impossible to find training images that are large enough to yield representative statistics.

To a certain degree, I comply with these concerns. A good example shows up naturally when one looks at the multiple-points statistics extracted from a training

image. As soon as the image presents some degree of complexity, most of the data events have only one single replicate (Hu and Chugunova, 2008). The question of whether this can still be called statistics remains open, but concerning realism of the resulting parameters fields, the efficiency of the method is unquestioned. My view is that statistical soundness is subordinate to realism in empirical sciences such as hydrogeology, because the overwhelming complexity of the processes and the interpretative character of the available data mostly produce particular cases that may have little statistical significance. I do not mean that the statistical parameters inferred from the data are irrelevant and should be systematically discarded. They should be considered as an approximate clue and not accepted as a genuine model.

Mathematicians and geologists may have different views on this question, and it is in my opinion a major cause of the widespread disinterest of hydrogeologists for stochastic methods (Christakos, 2004; Renard, 2007).

If realism is the most important criterion for validating a model, why at all shall we use mathematical and statistical methods for generating images of natural processes? These methods are designed to give objective answers to clearly defined problems, and not to produce something as shapeless as "realism". Why not drawing these images by hand instead? I think hand-drawn images are not bad and should be considered when possible, and this is often the case in practice. One can argue that it does not allow estimating uncertainty. This is not entirely true because it is common practice to consider multiple geological scenarios, which is a kind of multiple-realization approach. Hand-drawn images are subjective, but so are quantitative methods. In most stochastic models, subjective decisions are also taken whether to define prior estimations of parameter distributions, to define different facies, to delineate the limits of the modeled area or even to choose one stochastic method over another.

In 3D cases, when numerous conditioning data or indirect data must be integrated, when different sources of information have to be combined, or when image generation has to be automated (such as in the framework of an inverse problem), it is not feasible to draw the images by hand. Then, stochastic image generation methods are needed. The utility of such methods is clear. They generally help characterizing geological formations and are of great use for dealing with a variety of problems (De Marsily*, et al.*, 2005). Nevertheless, their mathematical and statistical soundness is not a guarantee that physical processes are correctly represented. Instead, I think statistical soundness is merely one of the prerequisites that may eventually lead to obtaining realistic images.

## 2. Perspectives

The main focus of this thesis is the generation of variables fields using stochastic techniques. Several geostatistical simulation algorithms were developed. Among them, the Direct Sampling has specific characteristics that can be explored beyond what has been done in this thesis.

The use of distances between data events has particular advantages. Some custom distances have been suggested in chapter 2, but more sophisticated ones can be used. A distance has been proposed that is independent of the mean value of the data event. One can imagine distances that are independent of other characteristics of the data event, such as an angle-invariant distance that would be computed by scanning the training image with data events rotated at various angles. A pattern containing a channel oriented NS could match a pattern containing a channel oriented EW. Similarly, a scale-independent distance would consist in scanning the training image with data events presenting different affinity ratios.

The possibility to perform multiple-points multivariate simulations can be further investigated. It was shown that multiple-points correlations can be accounted for and reproduced. But what can be characterized by a multiple-points correlation? For example, if the correlation between hydraulic conductivity and hydraulic head can be entirely characterized (together with boundary conditions), is it possible to co-simulate them and have a proxy solution to a flow problem?

The simulation of continuous properties could also open research directions for inverse problem coupled with multiple-points statistics, because it makes it possible to compute the derivative of the simulated values with respect to an objective function, which is impractical when simulating facies.

The nature of the algorithm makes it easy to parallelize on shared memory architectures (this has been discussed in chapter 2). This feature can be further exploited by using graphical processing units (GPUs), which have a much greater computational potential than CPUs for a lesser hardware purchase cost (NVIDIA Corporation, 2009).

The question of the possible applications of the generated images has only been discussed marginally. One aspect of future work to be carried on is to apply these methods on real-world hydrogeological problems. This would allow defining what are the key advantages and weaknesses of the methods developed. Furthermore, new fields of application can be envisioned. For example, the simulation of financial indices is a continuous multivariate problem that can be addressed with Direct Sampling. The amount of historical data is by far large enough to obtain good

training images or training datasets. Numerous other 1D applications are possible in very diverse domains, such as simulating sound frequencies to generate synthetic music or speech. The reconstruction algorithm could be used to restore damaged photographs, movies or paintings. The super-resolution method could be used in digital imaging software. All these applications could not be tested in the timeframe of this PhD project, but may be worth considering for future work.

## 3. References

Christakos, G. (2004), *A sociological approach to the state of stochastic hydrogeology*, Stoch Envir Res and Risk Ass, 18, 274–277.

De Marsily, G., Delay, F., Gonçalvès, J., Renard, P., Teles, V., and Violette, S. (2005), *Dealing with spatial heterogenity*, Hydrogeology Journal, 13, 1, 161-183.

Hu, L., and Chugunova, T. (2008), *Multiple-Point Geostatistics for Modeling Subsurface Heterogeneity: a Comprehensive Review*, Water Resour. Res., 44, W11413.

Journel, A., and Zhang, T. (2006), *The Necessity of a Multiple-Point Prior Model*, Mathematical Geology, 38, 5, 591-610.

NVIDIA Corporation (2009), *NVIDIA CUDA Programming Guide*, NVIDIA Corporation.

Renard, P. (2007), *Stochastic hydrogeology: what professionals really need?*, Groundwater, Manuscript submitted.

# Appendixes

# Appendix A

## Truncated plurigaussian simulations to characterize aquifer heterogeneity*

* This chapter was published as:
Mariethoz, G., P. Renard, F. Cornaton, and O. Jaquet. (2009). Truncated plurigaussian simulations to characterize aquifer heterogeneity. Ground Water Vol. 47, No.1 p. 13-24.

**Abstract** Integrating geological concepts, such as relative positions and proportions of the different lithofacies, is of highest importance in order to render realistic geological patterns. The truncated plurigaussian simulation method provides a way of using both local and conceptual geological information to infer the distributions of the facies and then those of hydraulic parameters. The method (Le Loc'h and Galli, 1994) is based on the idea of truncating at least two underlying multi-Gaussian simulations in order to create maps of categorical variable. In this manuscript we show how this technique can be used to assess contaminant migration in highly heterogeneous media. We illustrate its application on the biggest contaminated site of Switzerland. It consists of a contaminant plume located in the lower fresh water Molasse on the western Swiss Plateau. The highly heterogeneous character of this formation calls for efficient stochastic methods in order to characterize transport processes.

## 1. Introduction

The importance of geostatistical characterization of complex geology has increased significantly over the last two decades. In this framework, the usual continuous and multi-Gaussian methods (Matheron, 1965) have shown that they do not allow to model a sufficiently wide range of connectivity patterns for the high (or low) permeable structures (Journel and Alabert, 1990; Kerrou*, et al.*, 2008; Renard, 2007; Renard*, et al.*, 2005; Zinn and Harvey, 2003). An alternative is to use a two-step approach in which, first the geological facies are modeled and second, they are populated with heterogeneous hydraulic and transport parameters. This approach is flexible and allows modeling structures at different scales.

When geological facies can be identified from field observations, a large number of methods can be used to simulate such categorical variables (see review in De Marsily*, et al.*, 2005; see review in Koltermann and Gorelick, 1996). Among the first were the sequential indicator (Journel and Isaaks, 1984), the Boolean (Haldorsen and Chang, 1986) and the truncated Gaussian (Matheron*, et al.*, 1987) methods. The use of sequential indicator simulation is progressively fading, mainly because it fails to represent complex geological structures. Furthermore, it leads to consistency problems as the generated simulations present multivariate distributions that are implementation-dependent (Emery, 2005). The Boolean approach produces realistic geometries (Deutsch and Tran, 2002; Jussel*, et al.*, 1994; Scheibe and Freyberg, 1995). Geological processes such as deposition and erosion can be included in the framework of a mixture of object based and pseudo-genetic algorithm (Cojan*, et al.*, 2004; Webb and Anderson, 1996). With respect to data conditioning with the Boolean model, an efficient iterative algorithm was proposed

by Lantuéjoul (2002). However, difficulties remain in the estimation of the parameters in order to constrain the size, shape, and density of the simulated objects. The Markov chain approach (Carle and Fogg, 1997) is a powerful alternative which makes use of transition probabilities between the facies. It has been applied to a wide variety of situations (Weissmann and Fogg, 1999). More recently, the use of support Vector Machines has been proposed (Kanevski*, et al.*, 2002; Wohlberg*, et al.*, 2006). The technique can delineate facies using regression techniques, but it does not allow sampling the probability space as it only produces a single facies realization. Multiple-points geostatistics (Caers and Zhang, 2004; Feyen and Caers, 2006; Strebelle, 2002) is very promising. The technique offers more flexibility than the Boolean approach and facilitates conditioning. However, it is computationally demanding (especially in terms of memory requirements) for large 3D grids with more than 4 facies according to our experience.

In this paper, we investigate the applicability of the truncated plurigaussian method (Le Loc'h and Galli, 1994). The principle of the method is to simulate one or several continuous Gaussian fields and to truncate them in order to produce a categorical variable. To illustrate the concept, let us consider a single Gaussian (continuous) variable and a single truncation: if the simulated Gaussian variable is above the threshold then the point belongs to the facies 1, if it is below the threshold then the point belongs to the other facies. This idea was investigated by Isaaks (1984) and was further developed by Matheron et al. (1987). The main interest of the truncated plurigaussian method is that it allows integrating a geological conceptual model (using a *lithotype rule*) within the framework of a mathematically consistent stochastic model while remaining tractable for large and high resolution 3D grids. It also presents the advantage that conditioning can be achieved in the presence of substantial datasets with acceptable run-times. The technique is for the moment mostly used in the petroleum (Remacre and Zapparolli, 2003) and mining industries (Fontaine and Beucher, 2006) where it has been widely validated. To the best of our knowledge, it has not yet been applied to hydrogeology. To demonstrate its applicability, we focus on the case study of the Kölliken contaminated site. This site was chosen because of the existence of an extensive data set (245 borehole logs with geological descriptions, existence of a tunnel, continuous measurements of contamination in observation wells, hydraulic tests, etc).

## 2. Truncated plurigaussian simulations

The mathematical theory underlying the truncated gaussian (Matheron*, et al.*, 1987) and the truncated plurigaussian (Le Loc'h and Galli, 1994) methods is described in detail in the book of Armstrong et al. (2003) or in the paper of Emery (2005). Therefore, we will emphasize here only the main aspects of the method and present them in an intuitive manner without entering into its mathematical formalism. The principle of the method is to generate two (or more) Gaussian fields, using standard multi-Gaussian techniques, and then to truncate them in order to produce a map of discrete values representing the lithotypes (Le Loc'h and Galli, 1994). The statistical inference of the variograms of the underlying Gaussian fields and their conditional generation will be discussed later. Let us first focus on an illustration of the truncation procedure and on the flexibility that it provides to the modeler. Figure 1 shows two Gaussian random fields: G1 and G2. In that case, G1 has a gaussian variogram model, while G2 has a spherical one and presents an East-West anisotropy. Both fields have a 0 mean and a variance of 1. These two fields are truncated to create 4 lithofacies. Because the relations between the facies can be different depending on the type of geology, the truncation technique has to be flexible. The approach proposed by Le Loc'h and Galli (1994) for that purpose is to define the relations between the facies in a diagram called the *lithotype rule*. Three examples of lithotype rules (A, B, and C) are shown in Figure 1. In these diagrams, the two axes correspond to the values of the underlying multi-Gaussian fields (G1 and G2), and the grey codes correspond to the domain of the different lithotypes. Figure 1 shows the application of the different lithotype rules. With the truncation rule A, G2 is truncated with two thresholds, defining the sand, silt and clay facies. Another threshold has been placed along G1, delimiting the basalt facies. The result (Figure 1a) is that whenever G1 has a low value, the basalt facies is present. At locations with a higher value of G1, sand, silt or clay is present according to the value of G2. Because G1 has a Gaussian covariance, the boundary between the basalt facies and the other facies has a smoother shape than the boundaries between the sands, silts and clay which is controlled by the spherical model of variogram used to generate G2. The relations and contacts between the facies are imposed by the lithotype rule. In this example, silt can be in contact with all other facies, but sand and clay are not allowed to appear next to each other. On the contrary, basalt can cut all other facies. Note that the surface areas of the different facies in the lithotype rule do not correspond to their respective proportions in the simulation because the underlying continuous variables are not uniformly distributed (they are

Gaussian). The control of the proportion in a given simulation requires therefore to compute precisely the value of the threshold (see details in Armstrong*, et al.*, 2003). Lithotype rule B has 3 facies that are in a fixed order, silt is a transition facies between sand and clay. One could think that such a result could be obtained by truncating a single Gaussian function. But by looking carefully at the spatial structure of the clay patches (figure 1b), it becomes visible that they are influenced by the spatial structure of both G1 and G2, with some smooth clay patches and other more irregular ones. Lithotype rule C shows that a facies can also be defined by discontinuous zones in the lithotype rule, generating complex effects (Figure 1c).

The choice of a lithotype rule is therefore a major step of the methodology. In practice, transition probabilities calculated from borehole logs provide good indications on which facies can and cannot be in contact. However, this is not sufficient since it is restricted to the vertical transitions. Therefore the lithotype rule is usually based on both the analysis of the borehole logs and on a geological conceptual model.

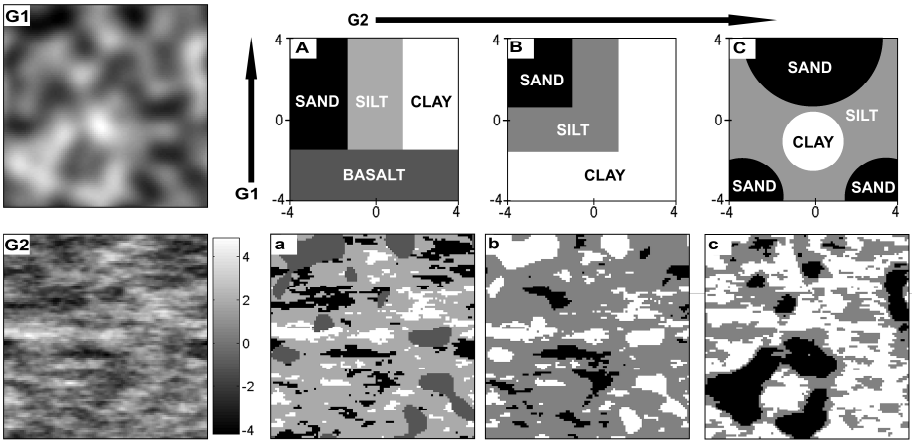



**Fig 1** G1 and G2: Underlying Gaussian fields (100x100 cells). A: Lithotype rule with 4 facies and a: corresponding simulation. B: Lithotype rule with 3 facies, all influenced by both G1 and G2 and b: corresponding simulation. C: Lithotype rule with facies that are defined by discontinuous zones instead of thresholds and c: corresponding simulation. Note that the areas of the different facies in the lithotype rules do not correspond to their respective proportions in the simulation because the underlying continuous variables are not uniformly distributed, but Gaussian.

To compute precisely the values of the threshold, one needs to define the relative proportions of the different lithofacies that will be simulated. These proportions are estimated by analyzing wells or outcrops data. Furthermore, in most practical cases, these proportions are not constant over the domain, but vary vertically and laterally because of the existence of trends in the geological processes. This non-stationarity is modeled by providing variable proportions over the domain. The lithotype rule is then locally updated by adjusting the values of the thresholds to match the target proportions while preserving the respective positions of the lithofacies.

An important feature of the plurigaussian technique is the inference of the variogram models for the underlying multi-Gaussian fields. Direct adjustment to the experimental variograms is not possible since the only available experimental variograms are the variograms of the indicator functions describing the lithofacies (one per lithofacies, plus all the bivariate combinations) while the two variograms needed for the model are the variograms of the underlying and continuous multi-Gaussian functions. The links between all these variograms are complex functions of the truncation process and of the conditioning to the lithofacies proportions. Therefore, the variogram inference is based on an inverse procedure in which the ranges of the variograms of the multi-Gaussian fields are adjusted iteratively through an inverse procedure. It consists in defining first the type and parameters of the initial variogram models, then these variograms are used to construct an unconditional plurigaussian simulation. One can then compute numerically the variograms of the indicators of the facies from the simulated field, and adjust the parameters of the variograms until an acceptable match is obtained between the experimental and the computed variograms (as described in figure 2). This is done automatically using a least-squares gradient-based minimization technique.

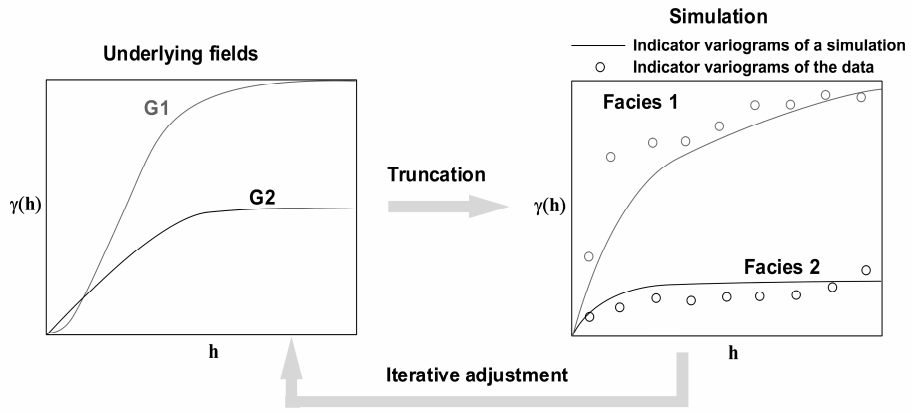



**Fig 2** The variogram models of the underlying Gaussian fields (left) are iteratively adjusted until the indicator variograms of the resulting truncated simulation matches the experimental indicator variograms of the field data (right).

The last step of the plurigaussian method that needs to be explained is the conditioning to borehole data. Again, because the method is based on the simulation of underlying multi-Gaussian fields and not on the simulation of the indicator variables, the conditioning cannot be direct. Two approaches are possible. The most rigorous is to impose local inequality constraints to the simulation of the multi-Gaussian fields. This can be achieved using the Gibbs sampler (Geman and Geman, 1984). This iterative algorithm was adapted to truncated Gaussian simulations by Freulon and Fouquet (1993) and is described in detail in Armstrong et al. (2003). The principle of the algorithm is to iteratively re-simulate a large number of times the Gaussian field until it reproduces both the structural model (variogram) and the constraints. Starting from an initial simulation, each pixel is resimulated accounting for the previously simulated values and accounting for the constraints. All simulated values that do not satisfy the constraints are rejected by the algorithm. After a certain number of iterations the process converges and honors all the required properties.

# 3. Application

## 3.1. Site description

The test site is the Kölliken waste landfill in central Switzerland. Between 1978 and 1985, 320'000 tons of special waste materials were buried in this ancient clay quarry. The waste consisted mostly of products of the chemical industry and incineration ashes. Due to economical reasons and improper legislation at that time, no impervious layer was installed to prevent leakage towards the underlying sandstone units. Moreover, a good drainage system was not implemented. As a result, Kölliken is today the biggest contaminated site in Switzerland. It has been investigated extensively and strong remediation measures have been applied (Abbaspour*, et al.*, 1998; Hug, 2004; , 2005). The main difficulty was the highly heterogeneous character of the site. The geological formations belong to the lower fresh water Molasse of the Swiss plateau. They consist of a succession of sandstones and marls corresponding to a setting of terrestrial deposition with meandering channels (Berger, 1985; Sommaruga, 1997). Given the amount of non-degradable pollutants, a pump-and-treat approach was first adopted. The remediation project was extended in 2003 by drilling a drainage tunnel along the southern side of the site, collecting the water downstream the landfill on 129 drainage wells at a depth of up to 20 m, combined with an on site water treatment plant. The purpose of this tunnel was to create a piezometric depression stopping further leakage. However, a small part of the plume is already far away downstream and cannot be recovered by pumping. This plume is now advancing in the Molasse formation and may reach the overlying alluvial aquifer. This aquifer supplies drinking water wells, the closest one being 4 kilometres downstream of the landfill. The limit between the alluvial aquifer and Molasse is a smooth but irregular surface of erosion. The motivation to use a high resolution geological model for the Kölliken site is the presence of thin, high permeable and well connected features in the Molasse that one can observe on outcrops. These structures are expected to control most of the fast contaminant migration. The aim of the research being methodological, only the part of the Molasse formation that is just below the erosion surface and that contains the part of the contaminant plume not captured by the drainage system has been selected for the model. The model extension (Figure 3) has a rectangular extension parallel to the cardinal directions. The geological data set consists of 245 boreholes logs and 219 measurements taken along the drainage tunnel. 9 cross sections have been made by integrating the data and the geological knowledge (Hug, 2005).

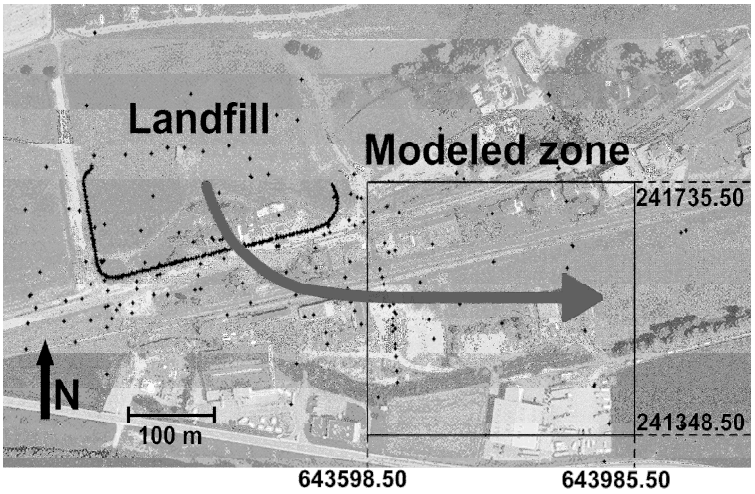



**Fig 3.**. Situation of the modelled zone and boreholes location (dots). The line of wells along the drainage tunnel is visible on the southern side of the landfill. Coordinates are in the CH1903 Swiss coordinate system.

## 3.2. Conceptual geological model

A thorough geological analysis of the site is essential in order to build a valid structural model. This is a major step because hard data (such as borehole logs) are usually insufficient to define the type of internal structure of the aquifer.

From a geological standpoint, Molasse is a thick Tertiary sedimentary body created by the detrital filling of a subsidence basin that was caused by the uplift of the Alps. With increasing paleo-distance from the Alps, one can find sediments ranging from very thick alluvial debris fans to deep marine turbidites. The total thickness of the Molasse formation can reach up to 5000 m on its southern side and is thinning up northwards. At the time of deposition, terrestrial debris arrived continuously from the Alps, generating more subsidence. Together with eustatic variations, this led to four different stages of marine and terrestrial deposits (Berger, 1985). These four stages are classically described as UMM (first marine stage), USM (first terrestrial stage), OMM (second marine stage) and OSM (second terrestrial stage). The Kölliken site lies within the USM in which the sedimentary structures are a succession of sandstones and marls. Together with paleogeographical and stratigraphic information, the geological interpretation of the area depicts an alluvial plain with meandering rivers. Crevasse splays deposits intersect levees, and the channel belts are wandering through the alluvial plains

scattered with marshy spots (Keller, 1992; Keller*, et al.*, 1990). Very detailed core analyses have identified 42 different facies, some of them highly represented and others very scarcely. Such a level of detail cannot be handled within the plurigaussian simulation framework, and these facies have to be grouped in a set of major lithotypes. Grouping cannot be done according to hydraulic parameters only, as this would lead to group facies with very different types of geometry and connectivity. For example, thin clay layers of lacustrine sediments and thick floodplain deposits may have the same conductivities but the simulation of their spatial distribution must be made separately in order to reproduce these differences and spatial structures because they will have different impacts on flow and transport.

To define which facies need to be modeled, we first have to understand the geology and the architecture of the USM formation. The most active element of such a system is the *river channel* (Figure 5a) moving sideways by erosion and deposition on the outer and inner banks. The point bar, where the slow motion of water allows for deposition of the suspended load and bed load, is mostly made of coarse sediments. A vertical section through a point bar deposit exhibits a gradation from coarser sand at the base to finer at the top (Nichols, 1999).

Repeated deposition of sand close to the channel edge leads to the formation of a *levee*, a bank of sediment flanking the channel which is higher than the level of the floodplain. With time, the level of the bottom of the channel can be raised by sedimentation in the channel and the level of water becomes higher than the floodplain level. When the levee breaks, water loaded with sediment is carried out on the floodplain to form a *crevasse splay*, a low cone of sediment formed by water flowing through the breach in the bank and out in the floodplain. These sediments are a heterogeneous mix of coarse debris carried by the river and fine material taken from the levees.

Flooding is not limited to crevasse splays: when the volume of water being supplied to a particular section of the river exceeds the volume which can be contained within the levees, the river floods and over bank flow occurs beyond the limits of the channel. Most of the sediment carried out on the *floodplain* is suspended load which will be mainly clay- and silt-sized debris. As water leaves the channel, it loses velocity very quickly. This drop in velocity triggers the deposition of most of the suspended load as thin sheet over the floodplain. These sheets of sand and silt deposited during floods events are thicker near the channel bank because coarser suspended load is dumped quickly by the flood. In periodically flooded plains, large areas of standing water can develop and persist for years of months. These can be assimilated to small lakes and are represented by finely laminated

sediments. On most of the floodplain's area, the marly sediments are in contact with the atmosphere and plants start to colonize this free space. This kind of environment is characterized by thick layers of marls and paleosols.

Though it is not highly tectonized, the USM has endured deformation during the alpine orogenesis. The resulting small scale fracturation is not well documented for the Kölliken site but has been observed in the region. These fractures clearly influence the hydraulic conductivity of the system.

The Molasse formation has then been eroded at the end of the Tertiary period. The final deposition phase was Quaternary alluvial sediments filling the bottom of the valley.

### 3.3. Data analysis and grid construction

An important point that must be clarified before starting the stochastic modeling is that the main directions of continuity of the lithology are generally not horizontal and not constant in space because the sedimentological structures are often deformed by tectonic movements. Therefore, before starting the data analysis, one has to define what the horizontal level was at the time of deposition. This can be done by identifying in the borehole data base a reference horizon or by investigating the structural data available on the site and interpolating this reference horizon over the whole site even if it does not correspond to a unique lithology. Once this horizon is identified and constructed, the entire domain can be deformed by coordinate transforms in order to restore its initial state as described by Armstrong et al. (2003) (see figure 4).

The geological modeling, including variogram inferences and the generation of conditional plurigaussian simulations, is done in the deformed space. The back transform allows returning to the actual system of coordinates.

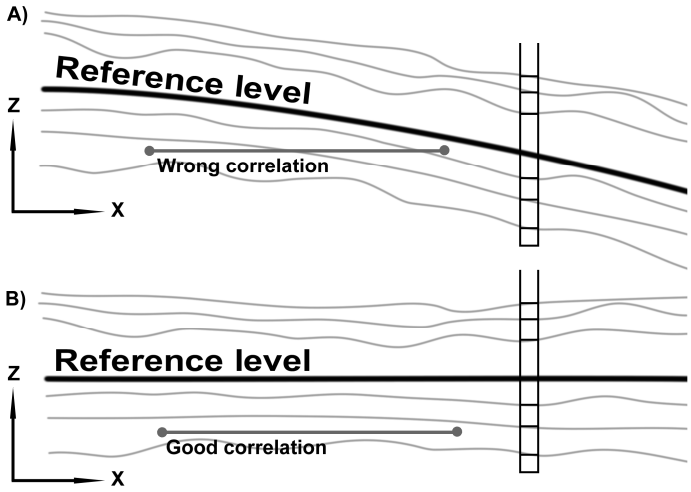



**Fig 4** A) Schematic sedimentary formation with reference level. Correlations inside a single layer are not horizontal. B) The same formation after flattening according to the reference level. Horizontal correlations are rendered possible to compute.

For the Kölliken site, a reference level has been constructed by drawing a line following the structures on all the geological sections, ensuring coherence between the sections, and interpolating a 2D surface from the data obtained from this interpretation. The interpolated surface is then used as a reference to modify the vertical component of the coordinate system in order to restore the data to their probable relative position at the time of deposition. A three dimensional regular grid is then built in this system of coordinate. The top of the model corresponds to the erosion surface between the Molasse and the alluvial aquifer. The bottom of the model is an arbitrary flat horizon at 340 m above sea level (about 120 m under the topographic surface). The volume is discretized into 2'984'725 blocks of 3 by 3 by 0.5 m.

## 3.4. Truncated plurigaussian simulations

Based on a preliminary geostatistical study performed by Thakur (2001) and considering the geological conceptual model described above, five facies have been considered: the channels (RG), the levee (UW), the crevasse splays (DFR), the paleosols in the flood plain (UPS), and the lacustrine deposits (LAK). We decided to

model the relations between the facies by the lithotype rule shown in figure 5b. This allows describing the lateral succession from channel to levee. The crevasse splay starts in a breach in the levee and can be in contact with all the other facies. Furthermore, crevasse splays have irregular boundaries as they surge in the alluvial plain and create deposits during sudden events. The first underlying multi-Gaussian function G1 is then modeled with a Gaussian covariance model (resulting in smooth boundaries between the channels and the associated structures) with initial ranges of 250 m in the EW direction, 80 m in the NS direction and 3m vertically. These ranges reflect the lateral and vertical extension of the channels. The second multi-Gaussian function G2 controls the position of the boundary between the crevasse splays and the other facies. On the contrary to G1, the ranges of G2 are longer on the NS direction (150 m) than in the EW direction (100 m) because the two structures are perpendicularly oriented. The initial vertical range is only 1.5 m as crevasse splays form very thin beds. G2 has an exponential variogram, which allows modeling an irregular boundary.

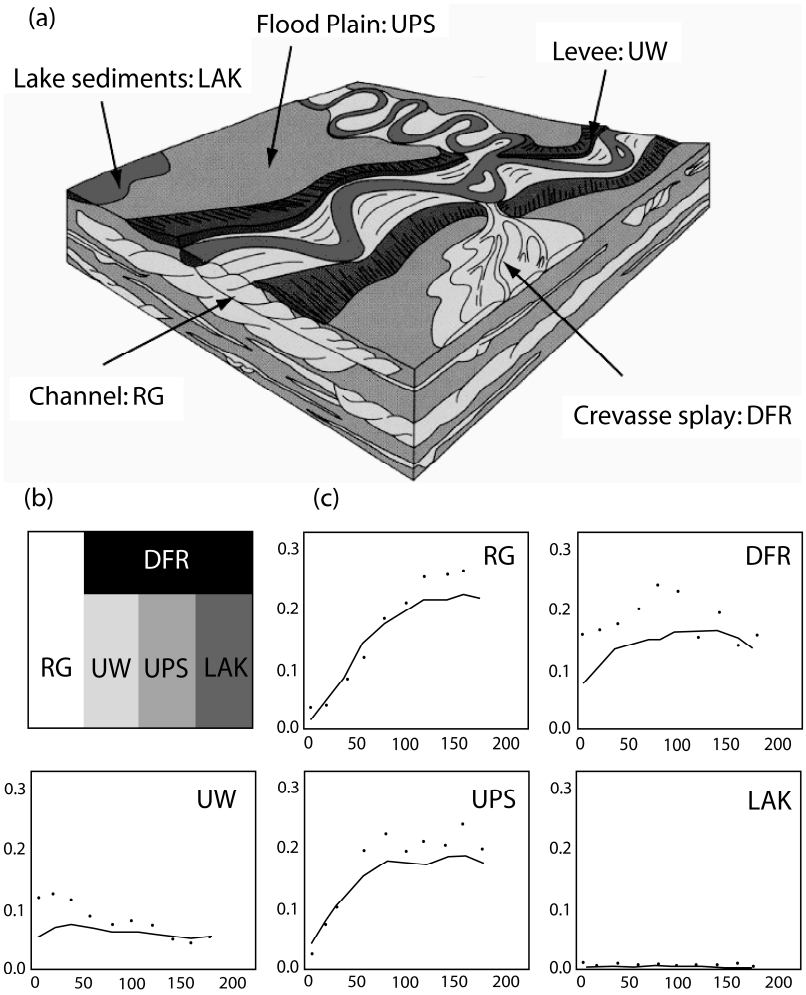



**Fig 5** (a) The conceptual model of the deposition environment and the associated lithotype rule. Modified from (Keller*, et al.*, 1990). (b) the lithotype rule, (c) fitted indicator variograms for the 5 lithofacies in the horizontal direction.

With the prescribed lithotype rule and 464 conditioning data, after having inferred the variograms by inversion (Figure 5b), 300 realizations are generated to attribute facies codes to all grid cells. As shown in Figure 5c, the variograms computed on the simulated fields (continuous lines) agree reasonably well with the

experimental variograms (dots). Similarly the proportions of the different facies in the simulations and those derived from the well data are almost identical. Then, the positions of the grid cells are back transformed to the actual system of coordinates. Figure 6 shows one resulting conditional realization. As expected, it is visually different from the conceptual geological model (figures 5a), but one has to remember that the conceptual geological model is used only to establish the lithotype rule. The truncated plurigaussian model does not aim at reproducing precisely the shape of the objects. It allows reproducing the conditioning data at the borehole locations, the relative position of the different facies through the lithotype rules, the statistical extension of the facies via the variograms (Fig 5b), and the relative proportions of the facies. To illustrate this last aspect of the method, a striking characteristic of the Kölliken site is the high proportion of crevasse splay facies (around 30%) and the quasi-absence of lacustrine deposits. This is a local feature different from the general conceptual geological model (figure 5a) but this is a feature that is reproduced by the model in all the simulations (figure 6).

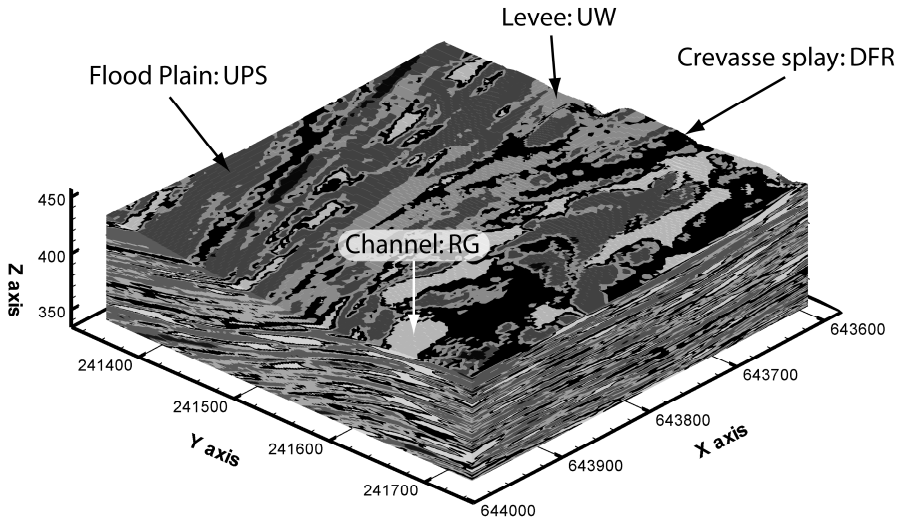



**Fig 6** One conditional realization generated by plurigaussian simulations. The simulation is shown after the back transform to the real coordinate system.

## 3.5. Flow and transport parameters

Among all the components present in the contaminant plume, Bromide was chosen for the simulation as it is a conservative tracer and a clear indicator of the contamination from the landfill. Its migration is modeled on each stochastic realization in transient state with a time-stepping scheme for a period of 15 years with the *Groundwater* finite element code (Cornaton, 2006). The flow field is assumed to remain in steady state.

The mean and variance of porosity and hydraulic conductivity for every facies have been measured in the laboratory on small plugs (3 cm high cylinders having a diameter of 3 cm) taken from boreholes cores (Dolliger, 1997; Keller*, et al.*, 1990). This data set has been collected in the same geological environment but not on the Kölliken site. To distribute the conductivities and porosities within the domain, we use this statistical information as well as non-conditional simulation because no data are available at the scale of the model elements on the Kölliken site.

In order to reproduce the correlation between porosity and permeability, porosity was modeled first as the combination of one random multi-Gaussian field for each facies. Note that the variogram used for the porosity simulation could not be inferred from on site data (because these data were not available) and was therefore estimated from geological knowledge by setting realistic ranges. Then the hydraulic conductivities ($K$) were estimated from the porosity with the Hagen-Poiseuille law:

$$K = \frac{n^3 \rho g}{b A_s^2 \mu}, \tag{1}$$

where $n$ is porosity [-], $b$ a formation factor (usually between 10 and 20), $A_s$ the specific contact surface between grains and water [$m^2/m^3$], $\mu$ the water viscosity fixed at 0.0027 [kg/m s], $\rho$ the water density fixed at 999.7 [$kg/m^3$] (for freshwater at 10°C) and $g$ the gravity acceleration, 9.81 [$m^2/s$]. The relation between porosity and hydraulic conductivity was obtained by adjusting the value of $A_s$ to optimize the fit of the data. The formation factor $b$ could be adjusted as well but because it constitutes a group with $A_s$ it is not possible to identify both of them separately so we decided to keep it fixed and equal to 20. A value of $A_s$ was obtained for each lithofacies, thus resulting in a relation (1) per lithofacies.

Furthermore, a white noise is added to represent the natural fluctuations that occur around the mean model. The variance of this noise has been estimated from the variance of the residuals between the measurements and the fitted Hagen-

Poiseuille law (Table 1). Two resulting hydraulic conductivity fields are shown in figure 7.

| **Facies** | **Nb of samples** | $A_s$ | **$\sigma^2 \log_{10} K$ of residuals** | **Mean $\log_{10} K$** | **$\sigma \log_{10} K$** | **Mean** $n$ | **Variance** $n$ |
|---|---|---|---|---|---|---|---|
| RG | 35 | 44000 | 1.22 | -5.95 | 1.46 | 0.209 | 0.003 |
| DFR | 23 | 166100 | 1.51 | -7.68 | 1.70 | 0.154 | 0.005 |
| UW | 21 | 700000 | 1.11 | -9.10 | 1.21 | 0.135 | 0.003 |
| UPS/ LAK | 3 | 1200000 | 0.55 | -9.56 | 0.38 | 0.112 | 0.003 |

**Table 1** Summary of the parameters used for the relationship between porosity and hydraulic conductivity. Hydraulic conductivity is expressed in [m/s].

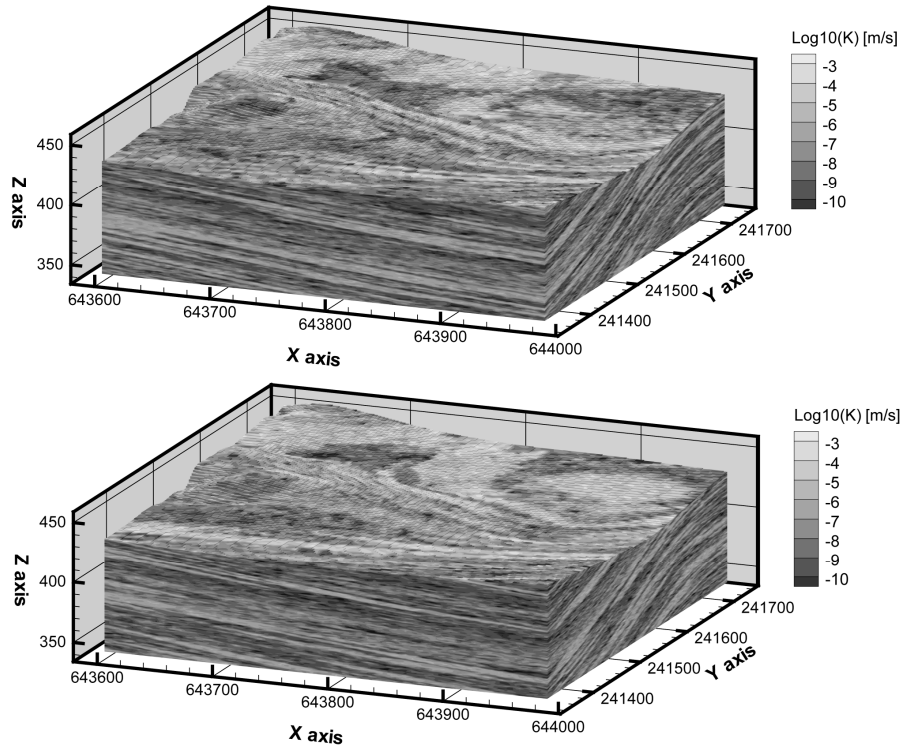


**Fig 7** The hydraulic conductivity field in log scale for two different realizations. While the general structure is similar, both realizations are clearly different, and lead to different contamination results.

In terms of transport parameters, the main dispersive process at the scale of the model is believed to be represented by the heterogeneity of the geological structure, therefore we set the longitudinal dispersivity coefficient to a value of 10 m and the transversal dispersivity to 1m. Even though these values could be lower, they significantly reduce the risk of numerical errors in the flow and transport simulation.

## 3.6. Initial and boundary conditions

The initial distribution of the Bromide concentrations is estimated by kriging 36 values measured during spring 2005 and ranging from 0.02 to 0.24 mg/l. This procedure is not optimal as the kriged field is not conditional on the geology and not constrained by the physics of solute transport. For example, it is highly probable that the contamination has not entered the low permeability formations and this is not accounted for. Furthermore, the uncertainty on this initial field is not evaluated while the variograms resulting from such sparse dataset are uncertain. All these limitations would have to be overcome in the case of an application with real practical implications.

As the limits of the model do not coincide with hydrogeological limits, it is not possible to prescribe boundary conditions corresponding to real physical boundaries. Nevertheless, previous regional models indicate that a bidirectional flow takes place on the vertical direction. This flow is driven by two nested flow systems. The first is local. It is caused by rainwater infiltrating on the escarpments and emerging in the bottom of the Kölliken valley in the modelled zone. It causes locally an upward flow component. The other system takes place at a bigger scale. A deep karstified calcareous bank underlying the USM drains the area and causes a general downward flow component. A water divide surface at depth within the Molasse separates the two systems. To represent this complex hydrogeological situation, it was chosen to impose head conditions on all sides of the model and to use a series of 1D Hermite polynomial interpolations along the vertical boundaries to force bidirectional flow. Again, in order to limit the complexity of the case study, we do not account for the transient variations of the head at the boundaries of the domain or the uncertainty on the boundary conditions.

The boundary conditions of the transport problem are kept very simple. The drainage tunnel, that lies downstream of the landfill and which was built in 2001, creates a piezometric depression capturing now all the leaking contamination. The consequence is that no new contaminant is arriving in the modelled area and this is why a zero concentration is prescribed to all inflowing zones of the system.

## 3.7. Model calibration

When simulating transport with the values of porosity and hydraulic conductivity described earlier, the plume migration was much slower than observed in the field. This difference is attributed mainly to the presence of small scale fractures which are not accounted for in the laboratory measurements on small plugs (sampling bias). To reproduce the mean velocity of the plume, the hydraulic conductivity needs to be increased. This observation is in agreement with previous description of the so-called scale effect in permeability (Clauser, 1992; Kiraly, 1988; Schulze-Makuch and Cherkauer, 1998; Zlotnik*, et al.*, 2000). We note that in addition to the sampling bias, there is an upscaling effect (Neuman and Di Federico, 2003; Renard and de Marsily, 1997) because the rock sample size is smaller than the grid blocks used in the model. This effect tends to reduce the variance and increase slightly the geometric mean of the distribution of the hydraulic conductivities in 3D, if we assume a classical multi-Gaussian model. This is however not sufficient to explain the discrepancy between the observed and modeled plume velocity. Therefore, we interpret this effect as due to the presence of small scale fractures within the USM formation which increase significantly the conductivity, but not the porosity. If all facies were made of consolidated rock, the fractures would have a uniformly distributed aperture. Therefore, the hydraulic conductivity of the fractures could just be added to the one of the matrix. Here, we made the hypothesis that the fractures are more open in consolidated sandstone than in clay or marls and, therefore, this effect increases the permeability differences between these facies. This can be rendered by multiplying all permeabilities by a given factor which has been adjusted by trial and error. A factor of 10 gave the best match between measured and calculated plume velocities. Note that an alternative approach could have been to reduce the effective porosity, however this would require a bias in the sampled porosity. We do not have sufficient data to support this hypothesis, and therefore we decided to use the simplest explanation compatible with our observations. Two of the resulting hydraulic conductivity fields are displayed in Figure 7.

## 3.8. Simulation results

The result of this procedure is a dataset of 300 realizations of concentration fields varying as a function of space and time. Overall, the plume is following the flow field and exits progressively the studied zone. The amount of contaminant leaving the area can thus be compared to the falling limb of a breakthrough curve. Figure 8 shows two vertical cross-sections through the domain for a particular realization at three time steps. From these raw results different statistical measures can be estimated, such as maps of mean concentration, probability maps (e.g. probability of having a concentration higher than a given threshold during a certain period of time) or a statistical distribution of global contamination fluxes.

Here, we present only the breakthrough curves, i.e. the history of the contaminant flux through the surface bounding the model on its eastern side (Figure 9). After a very fast drop of the flux during the first year, the simulations show an important variability with values ranging from 2 to 5 kg/y. These fluxes slowly decrease with time. For illustrating the importance of modeling heterogeneity from a practical point of view, the same calculations are made on a naïve homogeneous model with a constant homogeneous equivalent hydraulic conductivity. The value for this hydraulic conductivity has been estimated using the Landau-Lifshitz-Matheron conjecture for 3D isotropic media (Dagan, 1993; Ravalec-Dupin*, et al.*, 2000; Renard and de Marsily, 1997):

$$K_{eq} = \left\langle k^{1/3} \right\rangle^3 \tag{2}$$

where the brackets <> represent the average of the values of the local hydraulic conductivities $k$. We made this calculation on two data sets: the ensemble of all the hydraulic conductivities estimated from slug tests conducted on site, and the ensemble of our simulated (and calibrated) hydraulic conductivity fields. In the first case, we obtain $K_{eq}=3.4\ 10^{-6}$ m/s and in the second case, we obtain $K_{eq}=1.4\ 10^{-6}$ m/s. These two numbers are in good agreement and confirm that the factor of 10 used for the calibration of the hydraulic conductivities is reasonable. For the homogeneous model, we take the first value estimated from the slug tests as this is the value that one may have taken if the geological model would not have been constructed. The porosity is constant and equal to the arithmetic average of the local porosities: $n=0.17$. Results show that the homogeneous model underestimates the fluxes by a factor ranging from 1 to more than 2 orders of magnitude (Figure 9). Indeed, the homogeneous model does not account for the preferential flow paths formed by the

channels and, as a result, the progression of the plume is much slower than in the heterogeneous case.

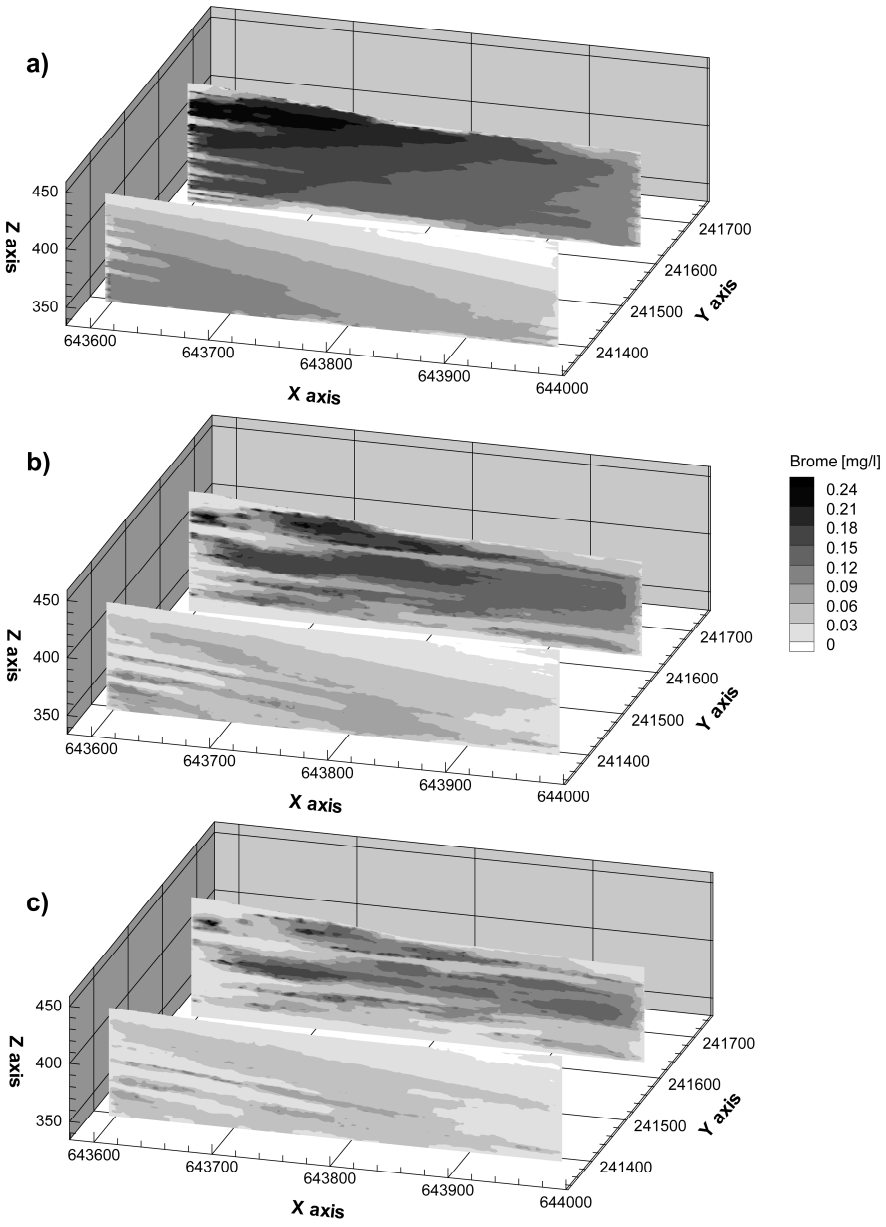


**Fig 8** a) The plume evolution after 0.5 years. b) 5 years. c) 10 years.

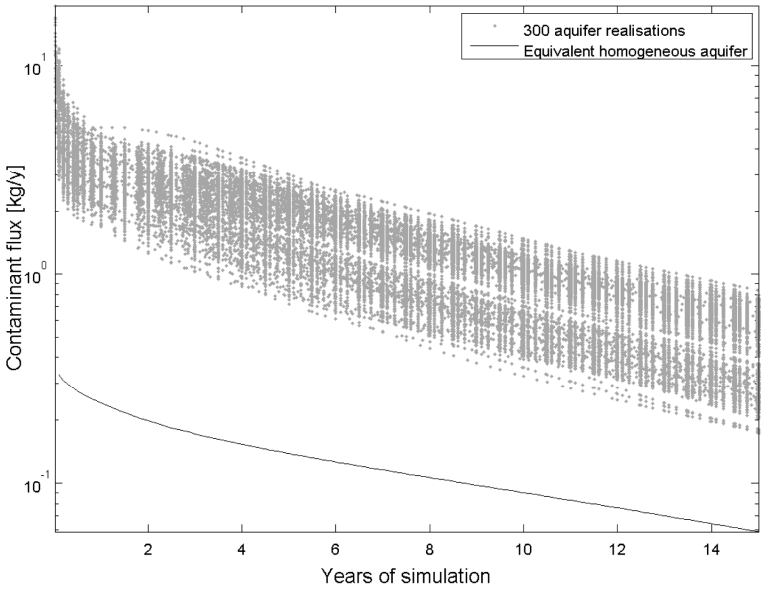



**Fig 9** Breakthrough on the eastern surface for 300 heterogeneous aquifer realizations.

## 4. Conclusion and discussion

Standard geostatistical models often suffer from a lack of geological realism due to restrictive assumptions such as for e.g. multi-Gaussianity that implies maximum entropy.. Like other techniques allowing to model lithofacies, the plurigaussian approach allows to circumvent this issue and model a wide range of connected or non connected (channels or barriers) geological structures that control flow and solute transport.

The main strength of the plurigaussian technique is that it allows incorporating a simple geological concept in the stochastic simulations. This is an important feature as in most applications, a detailed geological model is difficult to establish. The geological concept is not derived only from a statistical analysis of the well data but from a geological analysis of all the available information. It may also include the computation of probability of transitions between the facies along boreholes but the

interpretation will not be limited to that analysis. The geological interpretation is then formulated in terms of the lithotype rule indicating which facies can be in contact with which other facies. The mathematical theory allows modelling the variograms, incorporating facies proportion trends, and conditioning to local data within the framework of a well defined underlying statistical model. All these characteristics make the technique appealing for practical applications. In our view, an interesting aspect of the method is that the definition of the lithotype rule requires a close collaboration between the modeller and the field geologist. It forces a discussion between those two communities that work too often independently.

The main technical difficulty in applying the plurigaussian technique is the inference of the variogram models for the underlying multi-Gaussian functions. This is achieved through an iterative process that depends on the choice of an initial set of variogram parameters that may be difficult to identify. The procedure may fall into local optimums depending on the initial guess. In that case, it may be difficult to justify the use of a given variogram model. An important criterion to guide the choice of the variogram models is then the coherence with the geological interpretation, i.e. the variogram anisotropy and ranges can be guided by the geological expertise related to the size of the geological objects present on a given site (expected width of a channel for example). The other limitation of the method is that the lithotype rules are not defined with respect to given directions. Hence, it is not possible to impose, for example, that the levees are always on the side of the channel and not on the top. When a contact is defined in the lithotype rule, it may occur in all directions. This is however compensated by the possibility of imposing non stationarity on the proportions and changing for example the proportions of the different facies with depth.

## 5. References

Abbaspour, K. C., Schulin, R., Genuchen, M. T. v., and Schläppi, E. (1998), *Procedures for uncertainty analyses applied to a landfill leachate plume*, Ground Water, 36, 6, 874-883.

Armstrong, M., Galli, A. G., Loc'h, G. L., Geoffroy, F., and Eschard, R. (2003), *Plurigaussian Simulations in Geosciences*, Springer.

Berger, J.-P. (1985), *La transgression de la molasse marine supérieure (OMM) en Suisse occidentale*, Université de Fribourg, Fribourg.

Caers, J., and Zhang, T. (2004), *Multiple-point geostatistics: a quantitative vehicle for integrating geologic analogs into multiple reservoir models*, in *Integration of outcrop and modern analog data in reservoir models, AAPG memoir 80*, pp. 383-394, American Association of Petroleum Geologists.

Carle, S. F., and Fogg, G. E. (1997), *Modeling spatial variability with one and multi-dimensional continuous Markov chains*, Mathematical Geology, 7, 29, 891-918.

Clauser, C. (1992), *Permeability of Crystalline Rocks*, Eos, Transactions, American Geophysical Union, 73, 21, 233-240.

Cojan, I., Fouche, O., Lopez, S., and Rivoirard, J. (2004), *Process-based reservoir modelling in the example of meandering channel*, paper presented at Geostatistics Banff, Springer.

Cornaton, F. (2006), *GroundWater, A 3-D Ground water flow and transport finite element simulator*, Univ. Neuchâtel.

Dagan, G. (1993), *High-order correction of effective permeability of heterogeneous isotropic formations of lognormal conductivity distribution*, Transport in Porous Media, 12, 3, 279-290.

De Marsily, G., Delay, F., Gonçalvès, J., Renard, P., Teles, V., and Violette, S. (2005), *Dealing with spatial heterogenity*, Hydrogeology Journal, 13, 1, 161-183.

Deutsch, C., and Tran, T. (2002), *FLUVSIM: a program for object-based stochastic modeling of fluvial depositional systems*, Computers & Geosciences, 2002, 28, 525-535.

Dolliger, J. (1997), *Geologie und Hydrogeologie der Unteren Süsswassermolasse im SBB-Grauholztunnel bei Bern*, Bundesamt für Umwelt, Wald und Landschaft, Landeshydrologie und -geologie.

Emery, X. (2005), *Properties and limitations of sequential indicator simulation*, Stochastic Environmental Research and Risk Assessment, 6, 18, 414-424.

Feyen, L., and Caers, J. (2006), *Quantifying geological uncertainty for flow and transport modelling in multi-modal heterogenesous formations*, Advances in Water Resources 29, 6, 912-929.

Fontaine, L., and Beucher, H. (2006), *Simulation of the Muyumkum uranium roll front deposit by using truncated plurigaussian method*, paper presented at 6th international mining geology conference, Darwin, 21-23 august 2006.

Freulon, X., and Fouquet, C. (1993), *Conditioning a Gaussian model with inequalities*, in *Geostat Troia '92*, pp. 201-212, Kluwer.

Geman, S., and Geman, D. (1984), *Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images*, IEE Trans. Pattern Anal. and Mach Intel., 6, 6, 721-741.

Haldorsen, H. H., and Chang, D. M. (1986), *Notes on stochastic shales from outcrop to simulation models*, in *Reservoir characterization*, pp. 152-167, Academic.

Hug, R. (2004), *Entwicklung eines Sanierungskonzeptes für die Spitze der Schadstofffahne der Sondermülldeponie Kölliken (AG)*, Neuchâtel, CHYN.

Hug, R. (2005), *Hydrogeologische Untersuchungen im Abstrom der Sondermülldeponie Kölliken*, Bull. angew. Geol., 10, 1, 65.

Isaaks, E. (1984), *Indicator simulation: Application to the simulation of a high grade uranium mineralization*, in *Geostatistics for Natural Resources Characterization, Part 2*, pp. 1057-1069, D. Reidel Publishing Company.

Journel, A., and Alabert, F. (1990), *New Method for Reservoir Mapping*, Journal of Petroleum Technology, SPE paper 20781.

Journel, A., and Isaaks, E. (1984), *Conditional indicator simulation: Application to a Saskatchewan deposit*, Mathematical Geology, 16, 7, 685–718.

Jussel, P., Stauffer, F., and Dracos, T. (1994), *Transport modeling in heterogeneous aquifer: 1. Statistical description and numerical generation of gravel deposits*, Water Resour Res, 30, 6, 1803-1817.

Kanevski, M., Pozdnukhov, A., Canu, S., and Maignan, M. (2002), *Advanced Spatial Data Analysis and Modelling with Support Vector Machines*, International Journal on Fuzzy Systems, 4, 1, 606-615.

Keller, B. (1992), *Hydrogeologie des schweitzerischen Molasse-Beckens: Aktueller Wissensstand und weiterführende Betrachtungen*, Eclogae geol. helv., 85, 3, 611-651.

Keller, B., Bläsi, H.-R., Platt, N. H., Mozley, P. S., and Matter, A. (1990), *Technischer Bericht 90-41, Sedimentaere architektur der distalen unteren Süsswassermolasse und ihre beziehung zur diagenese und den Petrophysik. Eigenschaften am Beispiel der Bohrungen Langenthal*, NAGRA.

Kerrou, J., Renard, P., Hendricks-Franssen, H.-J., and Lunati, I. (2008), *Issues in characterizing heterogeneity and connectivity in non-multi-Gaussian media*, Advances in Water Resources, 31, 1, 147-159.

Kiraly, L. (1988), *Large scale 3-D groundwater flow modelling in highly heterogeneous geologic medium*, in *Groundwater Flow and Quality Modelling*, pp. 761- 775

Koltermann, C., and Gorelick, S. (1996), *Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches*, Water Resour. Res., 32, 9, 2617-2658.

Lantuéjoul, C. (2002), *Geostatistical simulation. Models and algorithms.*, Springer.

Le Loc'h, G., and Galli, A. G. (1994), *Improvement in the truncated Gaussian method: combining several Gaussian functions*, paper presented at Ecmor 4, 4th European Conference on the Mathematics of Oil Recovery, Roros, Norway.

Matheron, G. (1965), *Les variables régionalisées et leur estimation*, Masson.

Matheron, G., Beucher, H., Galli, A., Guérillot, D., and Ravenne, C. (1987), *Conditional simulation of the geometry of fluvio-deltaic reservoirs*, paper presented at 62nd Annual Technical Conference and Exhibition of the Society of petroleum Engineers, SPE Paper 16753, Dallas.

Neuman, S. P., and Di Federico, V. (2003), *Multifaceted nature of hydrogeologic scaling and its interpretation*, Reviews of geophysics, 42, 3, 1014.

Nichols, G. (1999), *Sedimentology and Stratigraphy*

Ravalec-Dupin, L. L., Noetinger, B., and Hu, L. Y. (2000), *The FFT moving average (FFT-MA) generator: An efficient numerical method for generating and conditionning Gaussian simulations*, Mathematical Geology, 32, 6, 701-723.

Remacre, A. Z., and Zapparolli, L. H. (2003), Application of the plurigaussian simulation technique in reproducing lithofacies with double anisotropy, in *Revista Brasileira de Geociências* edited, pp. 37-42.

Renard, P. (2007), *Stochastic hydrogeology: what professionals really need?*, Groundwater, Manuscript submitted.

Renard, P., and de Marsily, G. (1997), *Calculating equivalent permeability: A review*, Adv Water Res, 20, 5-6, 253-278.

Renard, P., Gómez-Hernández, J., and Ezzedine, S. (2005), *Characterisation of Porous and Fractured Media*, in *Encyclopedia of Hydrological Sciences*

Scheibe, T. D., and Freyberg, D. L. (1995), *Use of sedimentological information for geometric simulation of natural porous media structure*, Water Resour Res, 31, 12, 3259-3270.

Schulze-Makuch, D., and Cherkauer, D. (1998), *Variations in hydraulic conductivity with scale of measurement during aquifer tests in heterogeneous, porous carbonate rocks*, Hydrogeology Journal, 6, 204-215.

Sommaruga, A. (1997), *Geology of the central Jura and the molasse basin*, SNSN.

Strebelle, S. (2002), *Conditional simulation of complex geological structures using multiple point statistics.*, Mathematical Geology, 34, 1, 1-22.

Thakur, R. K. (2001), *Geostatistical simulations for 3D geological modelling of an industrial waste landfill*, Ecole des Mines de Paris, Fontainebleau (France).

Webb, E. K., and Anderson, M. P. (1996), *Simulation of preferential flow in three-dimensional, heterogeneous conductivity fields with realistic internal architecture*, Water Resour. Res., 32, 3, 533-546.

Weissmann, G. S., and Fogg, G. E. (1999), *Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a sequence stratigraphic framework*, Journal of Hydrology 226, 48-65.

Wohlberg, B., Tartakovski, D. M., and Guadagnini, A. (2006), *Subsurface characterization with support vector machines*, IEEE Transactions on Geoscience and Remote Sensing, 44, 1, 47-57.

Zinn, B., and Harvey, C. F. (2003), *When good statistical models of aquifer heterogeneity go bad: A comparison of flow, dispersion, and mass transfer in connected and multivariate Gaussian hydraulic conductivity fields*, Water Resour Res, 39, 3, 1051.

Zlotnik, V. A., Zurbuchen, B. R., Ptak, T., and Teutsch, G. (2000), *Support volume and scale effect in hydraulic conductivity: experimental aspects*, in *Theory, modelling, and field investigations in hydrogeology: A special volume in homor of Shlomo P. Neuman's 60th birthday*, pp. 215-231, Geological Society of America.

# Appendix B

## Reducing the impact of a desalination plant using stochastic modeling and optimization techniques*

* This chapter was published as:
Alcolea, A., P. Renard, G. Mariethoz, F. Bertone. (2009). Reducing the impact of a desalination plant using stochastic modeling and optimization techniques. Journal of Hydrology Vol. 365, Issue 3-4, p. 275-288.

**Abstract** Water is critical for economic growth in coastal areas. In this context, desalination has become an increasingly important technology over the last five decades. It often has environmental side effects, especially when the input water is pumped directly from the sea via intake pipelines. However, it is generally more efficient and cheaper to desalt brackish groundwater from beach wells rather than desalting seawater. Natural attenuation is also gained and hazards due to anthropogenic pollution of seawater are reduced. In order to minimize allocation and operational costs and impacts on groundwater resources, an optimum pumping network is required. Optimization techniques are often applied to this end. Because of aquifer heterogeneity, designing the optimum pumping network demands reliable characterizations of aquifer parameters. An optimum pumping network in a coastal aquifer in Oman, where a desalination plant currently pumps brackish groundwater at a rate of 1200 $m^3$/h for a freshwater production of 504 $m^3$/h (insufficient to satisfy the growing demand in the area) was designed using stochastic inverse modeling together with optimization techniques. The Monte Carlo analysis of 200 simulations of transmissivity and storage coefficient fields conditioned to the response to stresses of tidal fluctuation and three long term pumping tests was performed. These simulations are physically plausible and fit the available data well. Simulated transmissivity fields are used to design the optimum pumping configuration required to increase the current pumping rate to 9000 $m^3$/h, for a freshwater production of 3346 $m^3$/ h (more than six times larger than the existing one). For this task, new pumping wells need to be sited and their pumping rates defined. These unknowns are determined by a genetic algorithm that minimizes a function accounting for: (1) drilling, operational and maintenance costs, (2) target discharge and minimum drawdown (i.e., minimum aquifer vulnerability) and (3) technical feasibility of the solution. The performance of the optimum pumping network is compared to that of a synthetic, tradition-based hand-delineated design, where optimization is not performed. Results show that the combined use of stochastic inverse modeling and optimization techniques leads to minimum side effects (e.g., drawdowns in the area are reduced substantially) and to a significant reduction of allocation and operational costs.

## 1. Introduction

Approximately 44% of the world's population inhabits coastal areas, which represents more than the total world's population in 1950. Water is a critical factor for economic growth in these areas. Among the techniques devoted to providing freshwater in coastal areas, desalination has become increasingly important, especially during the last five decades. Actually, (Delyanns, 2003) found the first reference to desalination in the Bible (Exodus, 15:25; it reads of how Moses and the people of Israel came upon the waters of Merra, which were bitter: ''And he cried onto the Lord. And the Lord showed him a wood and he put it into the water and the water became sweet"). Being a simple technique, the use of desalination has spread

worldwide since 1950. Currently, growth is expected to be of 61% over a five-year period (from 39.9 million $m^3$/d at the beginning of 2006 to 64.3 million $m^3$/d in 2010 and 97.5 million $m^3$/d in 2015; GWI, 2006). However, the use of desalination is often controversial, for economic and environmental reasons. Disadvantages of desalination are the discharge of residuals, such as concentrated brine, back to the sea and an extremely high energy demand, which can add a significant contribution to greenhouse gas emissions. Often, input water is pumped directly from the sea through intake pipelines. This causes depletion of marine life (Dickie, 2007) due to impingement (i.e., death or injury due to contact with intake structures) and entrainment (i.e., marine life ‘‘sucked” by intake pipelines). Additional shortcomings are the large cost of off-shore constructions and the need for a prior filtration of seawater. These disadvantages prompted the development of new strategies. Desalting brackish groundwater from beach wells is usually more efficient (and therefore cheaper) than desalting seawater because (1) there is less suspended matter, so the filtration processes (and consequently the cost of necessary infrastructures) are reduced and (2) the pH of brackish groundwater is about 7 (8–9 for seawater), so no neutralization or additional chemical treatment is necessary. In addition, one gains natural attenuation and hazards due to anthropogenic pollution of seawater are reduced. Nowadays, desalted brackish water represents 24% of the worldwide freshwater production (GWI, 2005).

In this context, the environmental impacts and the cost of pumping brackish groundwater can be minimized by using optimization techniques(Abarca*, et al.*, 2006; Ahlfeld and Heidari, 1994; Ahlfeld and Mulligan, 2000; Gorelick, 1983; Wagner, 1995) (Ahlfeld and Heidari, 1994; Gorelick, 1983; Wagner, 1995). For coastal aquifers, these techniques have been applied either to optimize freshwater pumping networks (Cheng*, et al.*, 2000; Mantoglou, 2003) or to design remediation systems(Abarca*, et al.*, 2006; Ahlfeld and Heidari, 1994; Ahlfeld and Mulligan, 2000). Here, we use optimization techniques to design a pumping network for brackish groundwater. The aim is to achieve a target discharge while minimizing the environmental side effects and the demand of energy, thus minimizing the total cost of the solution. An important feature of the design is that it must be reliable regardless of the degree of aquifer heterogeneity and the corresponding uncertainty.

The need for a reliable design under uncertainty motivated the use of stochastic approaches, as opposed to single ‘best’ deterministic models (see reflections in Renard, 2007a; Renard, 2007b; Tarantola, 2005). For coastal aquifers, (Alcolea*, et al.*, 2007) integrated tidal fluctuation and injection tests in a stochastic model yielding a single ‘best’ estimation of the transmissivity and storage coefficient

fields. Since tides can be viewed as large-scale aquifer tests, they provide large-scale information on aquifer diffusivity and connectivity patterns. Hydraulic tests improve the identification of local connectivity (Carrera and Neuman, 1986b; Meier*, et al.*, 1998; Weiss and Smith, 1998) and allow resolving diffusivity into transmissivity and storage coefficient (Carrera and Neuman, 1986a; Rotting*, et al.*, 2006). In addition, geostatistical joint interpretation of data at all boreholes provides a continuous description of the connectivity structure (i.e., of diffusivity), rather than point values of effective diffusivity at a given set of boreholes (Li*, et al.*, 2007). Also the Monte Carlo type inverse framework is used frequently for the simulation of transmissivity fields. Yet, little attention has been paid to the joint simulation of transmissivity and storage coefficient (Hendricks-Franssen*, et al.*, 1999).

The objective of this paper is to demonstrate through a case study in Oman how stochastic modeling for aquifer characterization and non-linear optimization techniques can be applied to achieve a reliable design that can reduce both the environmental impacts of a desalination plant and minimize the costs of allocation and operation of the pumping system. The methodology suggested by (Alcolea*, et al.*, 2007) is extended to a Monte Carlo inverse framework and is used to characterize the spatial variability of both transmissivity and storage coefficient fields from the response to tidal fluctuation and to three long term pumping tests. In that manner, 200 equally likely simulations are conditioned to available data using the regularized pilot points method (Alcolea*, et al.*, 2006). The 200 simulations are physically plausible and fit the available data well. Next, transmissivity fields are used to determine the optimum pumping configuration using a genetic algorithm (Popov, 2005; Popov and Filipova, 2004) that minimizes a function accounting for: (1) the cost of allocation of wells and their maintenance, (2) the cost of electricity, which depends on drawdowns (i.e., minimum aquifer vulnerability) and (3) the technical feasibility of the solution, because only three different types of pump can be used.

The paper is organized as follows. First, the site is introduced. Second, the application of the characterization methodology is described and the results of the Monte Carlo analysis of the transmissivity and storage coefficient fields are displayed. The value of stochastic modeling is analyzed by comparing the outcomes of the conditional simulations with the 'single best' characterization obtained by conditional estimation. From that starting point, a description of the optimization procedure and the (single) optimum pumping configuration is presented. We then test the benefit of using optimization by comparing the performance of the optimum pumping configuration with the one obtained using a synthetic, tradition-based,

hand-delineated pumping network (i.e., no optimization is performed for the latter case). Last, some conclusions and recommendations are summarized.

## 2. Site description and conceptual model

The study area is located on the coast of Oman. The site is occupied by a desalination plant (Fig. 1). Brackish groundwater is pumped from beach wells at a rate of 1200 $m^3$/h, for a freshwater production of 504 $m^3$/h. The purpose of the underlying study is to design a pumping network that will allow increasing the overall pumping to 9000 $m_3$/h. This will provide a total freshwater production of 3346 $m^3$/h, sufficient to satisfy the growing demand of potable water.

The aquifer is made of sub-horizontal layers of early Palaeocene–Eocene fossiliferous limestone with interbedded conglomerates laying on top of a marl deposit. Surface observations show a primary porosity in the limestone due to the lack of compaction of the sediment. Limestone often presents karstic cavities, sometimes filled with sandy silt. Thus, it is expected that most of the secondary porosity is due to karstification processes, suggesting the presence of highly diffusive conduits in the area. This hypothesis was partially confirmed by a preliminary geophysical campaign (Fig. 1). Electrical resistivity data display the presence of structures with a clear orientation toward North. However, hydraulic conductivity of these bodies could not be inferred directly from resistivity data due to the presence of brackish water, both in the highly permeable (karstified zone) and in the clayey areas.

A preliminary drilling campaign revealed that groundwater inflows towards wells are located far underneath the groundwater table. In addition, short pumping tests after drilling provided relatively high estimates of transmissivity (between 0.01 and 0.3 $m^2$/s) not correlated with the saturated thickness. This clearly indicates that transmissivity is governed by the karstic features in depth but not by the rock matrix. Another important observation is that groundwater is brackish all over the site. This is confirmed by the set of measurements of electrical conductivity at available boreholes, which are very similar to that of local seawater. This is explained by the extremely low amount of recharge in the area. In fact, the interface between fresh and brackish water is located several kilometers inland. These two observations show that it is reasonable to, first, neglect 3D density effects and, second, to assume that the transmissivity does not depend on head variations. This allows us to use a linearized 2D approximation of the groundwater flow equation. As a consequence, the superposition principle applies during both the characterization and optimization

stages. This greatly accelerates all calculations. Yet, it is worth mentioning that the techniques used in this paper can also be applied (without too much modification) if the non linear groundwater flow equation is considered. In fact, we assume that the head variations are small relative to the aquifer thickness.

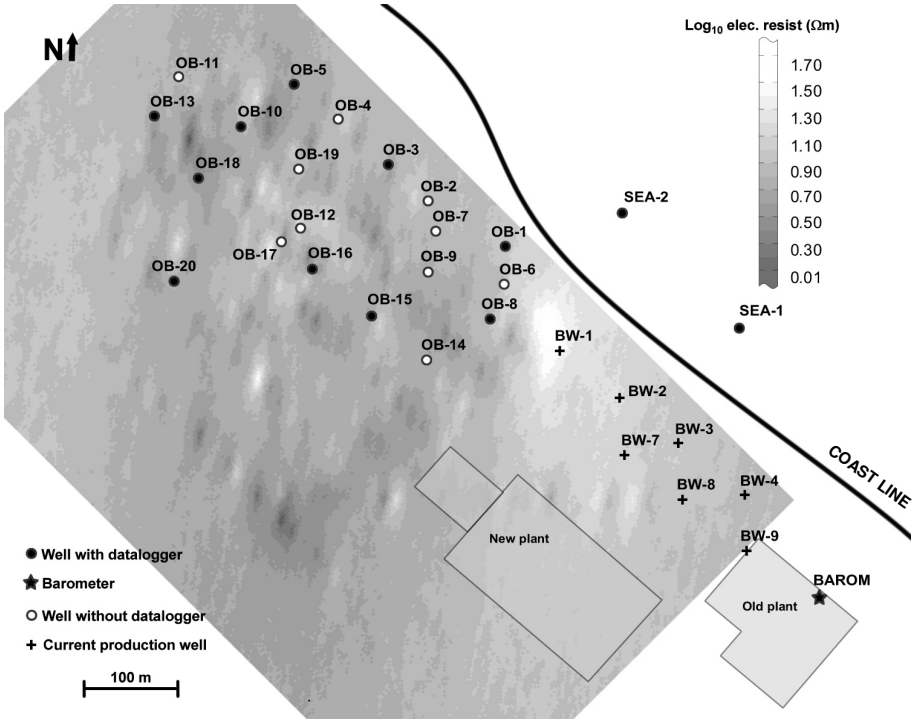



**Fig 1** Site description, observation (OB) and current production beach wells (BW). Some observation wells were automatically monitored and are depicted by dots. Two sensors (SEA-1 and SEA-2) were located at the sea-shore for measuring the sea level fluctuation. A barometer (depicted by a star) was located at the old desalination plant. The background image depicts the ensemble mean of 100 simulations of (vertically integrated) electrical resistivity arising from a preliminary geophysical campaign.

## 3. Hydraulic characterization of transmissivity and storage coefficient fields

The following is a detailed description of the characterization methodology, which includes data filtering, well testing and stochastic inversion. First, 200

simulations of the transmissivity and storage coefficient fields are conditioned to transmissivity, storage coefficient and head variation data (i.e., response to tidal fluctuation and to three long term pumping tests; Table 1) and their uncertainty is evaluated. The four head data sets are arranged in four independent flow problems, which are analyzed simultaneously. Second, for the sake of comparison, we also obtain a 'single best' solution by conditional estimation to the aforementioned data sets. This comparison illustrates the uncertainty overlooked by conditional estimation. Outcomes of these two sets are compared in terms of physical plausibility and fit to head variation data.

### 3.1. Available data

Absolute pressure ($p_{abs}$) was automatically recorded at the seashore (sensors SEA-1 and SEA-2) and at 10 boreholes (Fig. 1) every 30 s. Sensor SEA-2 served only as a backup in case of failure of SEA-1. First, very high frequency fluctuations of sea level due to waves and wind were filtered out as they are assumed not to propagate far away within the aquifer due to dampening effects and because the aquifer works as a high pass filter. A moving average algorithm was used to that end. Next, filtered measurements were transformed into relative pressures by subtracting the barometric pressure ($p_{rel} = p_{abs} - p_{bar}$), monitored with the same frequency at sensor BAROM. Next, relative pressures were transformed into pressure heads ($p_{rel}/\gamma$, where $\gamma$ is the specific weight of groundwater). Average specific weight at available boreholes was 1021 kg/m$^3$, very similar to that of local seawater (i.e., the entire study area has already been intruded by seawater). Finally, heads are calculated as the sum of pressure heads and sensor elevation. This is obtained as the difference between the elevation of a reference point at the well (e.g., top of casing) and the sum of the absolute pressure and the groundwater depth (dipped manually) at a given time in absence of pumping. It is good practice to calculate the sensor elevation at different times, corresponding to low, mean and high tide to ensure unbiasedness and statistical coherency of the methodology.

| Data in response to | Measurement period (in days) | Monitored wells OB- (Fig. 1) |
|---|---|---|
| Tidal fluctuation | 7 | 1, 3, 5, 8, 10, 13, 15, 16, 18, 20 |
| Pumping test at OB-6 | 4 | 1, 3, 5, 6, 8, 9, 10, 13, 14, 15, 16, 18, 20 |
| Pumping test at OB-15 | 2 | 1, 3, 5, 8, 9, 10, 13, 14, 15, 16, 18, 20 |

| | | |
|---|---|---|
| Pumping test at OB-16 | 2.5 | 1, 3, 5, 8, 9, 10, 12, 13, 15, 16, 17, 18, 20 |
| Transmissivity | - | All |
| Storage coefficient | - | 1, 3, 4, 5, 9, 10, 12, 13, 14, 15, 16 |

**Table 1** Data sets available for calibration. Measurement period of pumping tests includes the recovery.

### 3.2. Working with head fluctuations

Tidal response is expressed in terms of variation with respect to natural heads. This simplifies the boundary and initial conditions of the stochastic model, as one needs to simulate only the head variations induced by sea level fluctuations, but neither the regional flows nor the existing pumping of the desalination plant. To this end, head measurements at every borehole were corrected by subtracting their mean value. This operation simply shifts the recorded signal towards the horizontal axis (i.e., zero head fluctuation). Thus, if the signal at a well is coherent, head variations are bounded by sea level fluctuations. Calculated heads at sensors SEA-1 and OB-1 are displayed in Fig. 2. The large amount of data (~246,000 at every borehole) demands prior filtering consisting of the selection of one measurement every 15 min. This makes the data set manageable while allowing the selected measurements to capture the temporal variability of heads (Fig. 2).

### 3.3. Analysis of the tidal response

Analysis of the tidal response measured during two months before the start of the pumping tests allows us to estimate point values of effective diffusivity ($D_{eff} = T/S$, being $T$ transmissivity and $S$ storage coefficient). It is worth mentioning that diffusivity values at monitored boreholes will not be used as hard data for conditioning the stochastic model. Yet, they are valuable for checking the plausibility of the simulations. We follow roughly the steps of the tidal response method (TRM hereinafter, Ferris, 1951; Hvorslev, 1951):

$$\Delta H^{i}_{well} = \Delta H^{i}_{sea} \exp\left( -\sqrt{\frac{\pi x^2}{t^{i}_{0} D^{i}_{eff}}} \right) \tag{1}$$

where $\Delta H_{sea}$ and $\Delta H_{well}$ are the amplitudes of head fluctuations at the sea and at a well at distance $x$ inland from the coast, respectively, and $t_0$ is the period of the sea level fluctuation. A multi-component analysis is carried out considering the main

harmonics '*i*' of the sea level fluctuation. Five main harmonics were considered, with dominance of the semi-diurnal lunar principal wave (M2). Corresponding amplitudes were identified by analyzing the Fourier spectrum of the sea level fluctuation measured at sensor SEA-1 (Fig. 3). Following the same procedure, the corresponding harmonics of measured signals at monitored wells were identified. Next, effective diffusivity for a given harmonic can be obtained from Eq. (1). Table 2 summarizes the geometric average of estimated effective diffusivities (at monitored wells) using TRM, by prior interpretation of short term pumping tests and by the stochastic model. The large variability of estimated effective diffusivities confirms the highly heterogeneous character of the aquifer. A key issue in the application of TRM is the uncertainty on the knowledge of the distance between the well and the seashore. In this study, we have assumed the mean surface of seashore as boundary. Yet, the (true) contact between sea and aquifer should be accounted for. A 3D analysis of temperature profiles at the seashore may help to locate that contact.

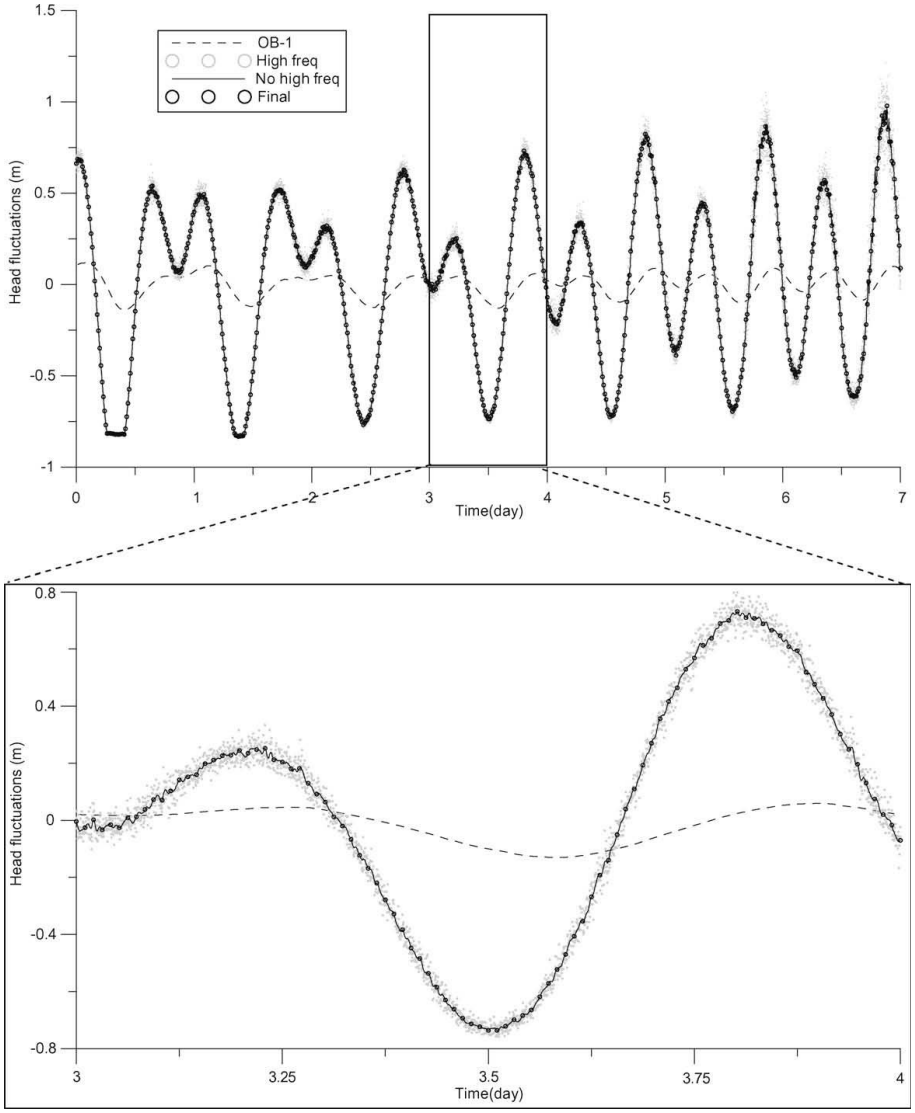



**Fig 2** Filtering the sea level fluctuation. Grey dots depict the measured sea level oscillation, containing high frequency fluctuations (i.e., due to waves and wind). The solid line is the result from a moving average filter for removing those undesired fluctuations. Finally, one measurement every 15 min is selected (black dots). Dashed line depicts the filtered head variation at observation well OB-1.

| | TRM | Pumping tests (HYTOOL) | | | Model |
|---|---|---|---|---|---|
| | $D_{eff}$ [$m^2/s$] | T [$m^2/s$] | S [-] | $D_{eff}$ [$m^2/s$] | $D_{eff}$ [$m^2/s$] |
| OB-1 | 0.33 | $1.6 \cdot 10^{-1}$ | 0.16 | 1.00 | 0.37 |
| OB-2 | - | $5.0 \cdot 10^{-2}$ | --- | --- | - |
| OB-3 | 0.95 | $9.2 \cdot 10^{-2}$ | 0.13 | $7.08 \cdot 10^{-1}$ | 0.97 |
| OB-4 | - | $3.3 \cdot 10^{-1}$ | $8.10 \cdot 10^{-2}$ | 4.07 | - |
| OB-5 | 1-95 | $1.7 \cdot 10^{-1}$ | 0.21 | $8.10 \cdot 10^{-1}$ | 1.28 |
| OB-6 | - | $1.6 \cdot 10^{-1}$ | --- | --- | - |
| OB-7 | - | $5.0 \cdot 10^{-2}$ | --- | --- | - |
| OB-8 | 3.35 | $1.8 \cdot 10^{-1}$ | --- | --- | 2.05 |
| OB-9 | - | $4.7 \cdot 10^{-2}$ | 0.18 | $2.61 \cdot 10^{-1}$ | - |
| OB-10 | 2.12 | $2.8 \cdot 10^{-1}$ | $3.80 \cdot 10^{-4}$ | 737.00 | 1.36 |
| OB-11 | - | $1.9 \cdot 10^{-1}$ | --- | --- | - |
| OB-12 | - | $5.2 \cdot 10^{-2}$ | 0.10 | $5.20 \cdot 10^{-1}$ | - |
| OB-13 | 0.75 | $2.2 \cdot 10^{-1}$ | 0.16 | 1.38 | 0.92 |
| OB-14 | - | $1.2 \cdot 10^{-1}$ | 0.10 | 1.20 | - |
| OB-15 | 0.52 | $2.9 \cdot 10^{-2}$ | $9.50 \cdot 10^{-2}$ | $3.05 \cdot 10^{-1}$ | 0.41 |
| OB-16 | 0.40 | $8.0 \cdot 10^{-2}$ | $2.40 \cdot 10^{-2}$ | 3.33 | 0.72 |
| OB-17 | - | $4.8 \cdot 10^{-2}$ | 0.34 | $1.41 \cdot 10^{-1}$ | - |
| OB-18 | 2.33 | $3.0 \cdot 10^{-1}$ | --- | --- | 1.58 |
| OB-19 | - | $1.0 \cdot 10^{-2}$ | --- | --- | - |
| OB-20 | 4.85 | $2.3 \cdot 10^{-1}$ | --- | --- | 1.73 |

**Table 2** Summary of effective diffusivities obtained by the tidal response method, by prior (conventional) interpretation of pumping tests and by the stochastic model (average of all simulations).

## 3.4. Prior interpretation of pumping tests

Standard interpretation of drawdown data (i.e., assuming homogeneity) allows us to obtain a prior estimation of the hydraulic parameters characterizing the aquifer. Unfortunately, hydraulic test data are not suitable for standard analysis due to the superposition of pumping and tidal effects (Chen and Jiao, 1999; Trefry and Johnston, 1998). The difficulty consists of estimating what should have been the natural heads during the pumping periods. Several alternatives can be used for separating these effects. Alcolea et al. (2007) used kriging with external drift. TRM estimated diffusivities can also be used to this end. None of these two methodologies yielded good results for available data sets, as confirmed by cross-validation.

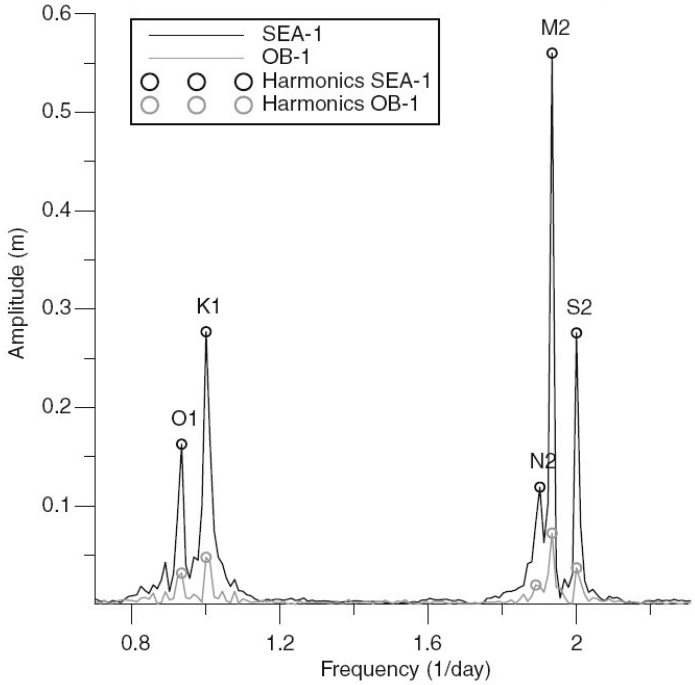



**Fig 3** Fourier spectrum of the sea level fluctuation and tidal response measured at sensors SEA-1 and OB-1, respectively. Five main harmonics are identified. From left to right, these are O1 (influence of lunar declinational diurnal wave), K1 (lunar– solar declinational diurnal), N2 (larger lunar elliptic semidiurnal), M2 (lunar semidiurnal) and S2 (solar semidiurnal). The tidal response at OB-1 is almost immediate, due to the proximity to the sea.

Instead, we used records at wells not affected by pumping and signal filtering algorithms. Natural heads $h_n^i(t)$ at a well '*i*' are estimated by:

(1) Selecting the signal at another (reference) well $h^{ref}(t)$, such that, first, signals $h^{ref}(t)$ and $h^i(t)$ are highly correlated in periods not affected by pumping and, second, $h^{ref}(t)$ is not affected by pumping.
(2) Iteratively correcting the amplitude and the phase of $h^{ref}(t)$:

$$h_n^i(t) = a\left[h^{ref}(t-\tau) - \left\langle h_{ref} \right\rangle\right] + b \qquad (2)$$

where $\tau$ is the time lag between both signals, $<h_{ref}>$ denotes the mean head at reference well, '*a*' is a dampening factor of amplitudes and '*b*' is a constant term that allows us to shift the heads. Parameters *a*, *b* and $\tau$ are estimated so that the differences between the reconstructed signal $h_n^i(t)$ and the measurements $h^i(t)$ are minimal in absence of pumping (i.e., we fit all measurements before and after pumping periods). Finally, drawdowns are calculated as the difference between natural heads and measured heads (See section ''Available data"). The pumping tests are interpreted by conventional analysis assuming homogeneous medium. This allows us to obtain prior estimates of transmissivity and storage coefficient at monitored wells. These values will condition the stochastic model. Prior interpretation of pumping tests was carried out using the open-source software HYTOOL (Renard, 2007a), a Matlab plug-in for the interpretation of hydraulic tests. This toolbox contains analytical solutions used to describe groundwater radial flow (e.g., Jacob, Boulton, Papadopoulos-Cooper, etc.), and functions for importing, displaying and fitting a model to available data. Table 2 summarizes the estimated values of transmissivity and storage coefficient.

## 3.5. Measurement errors

Acquisition errors were accounted for by assigning a standard deviation of 0.005 m to tidal response data (i.e., in absence of pumping). The filtering process described in previous sections was also accounted for by assigning a larger standard deviation of 0.01 m to the estimated drawdowns during pumping periods (Table 1). Prior information arising from the conventional interpretation of pumping tests was also uncertain. Main sources of uncertainty were the conceptual model (homogeneous) and those related to the short duration of the experiments. We assigned rather large error variances of 3 (in log scale) to the transmissivity and storage coefficient measurements used for conditioning the stochastic model (Table 2).

## 3.6. Spatio-temporal discretization

Two finite element codes have been used in this work: TRANSIN and GROUNDWATER. TRANSIN (Medina*, et al.*, 2000; Medina and Carrera, 2003) solves the geostatistical inverse groundwater flow and contaminant transport problem. It was used for characterizing aquifer parameters. GROUNDWATER

(Cornaton, 2006) provides an exhaustive mathematical representation of physical processes governing groundwater flow and contaminant transport. It is used at the optimization stage of this work. Both TRANSIN and GROUNDWATER allow solving groundwater flow under different assumptions. In this work, the codes are used to solve the 2D groundwater flow equation assuming that T is not affected by head changes. This is equivalent to assuming a standard 2D confined aquifer hypothesis (see section ‘‘Site description and conceptual model”). The finite element mesh is designed using the code 2DUMG (Bugeda, 1990). It honors the relevant geometric features (e.g., seashore, wells, etc.) and consists of 2585 nodes, arranged in 5074 triangular elements (Fig. 4). It is refined in the vicinity of the existing observation wells and close to the seashore. The element size increases as the mesh progresses outside the area encompassing the existing wells and/or inland. Temporal behavior is modeled with a forward in time finite differences scheme. The time step is constant (15 min, equal to the measurement frequency). Simulation periods are listed in Table 1.

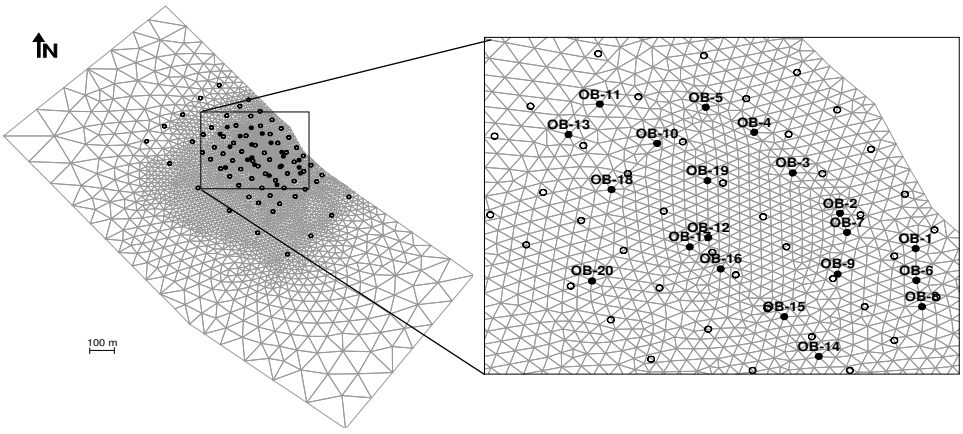



**Fig 4** Spatial discretization of the domain. The finite element mesh consists of 2585 nodes arranged in 5074 triangular elements. In the inset, existing (monitored) wells and pilot point locations are depicted by dots and circles, respectively.

## 3.7. Boundary and initial conditions

The methodology suggested by Alcolea et al. (2007) works with head fluctuations rather than with absolute heads. In this way, one needs to simulate only

head variations induced by tidal fluctuation or pumping, but neither the regional flow in the aquifer nor the existing pumping. Therefore, boundary and initial conditions are homogeneous (i.e., zero head variations and fluxes) and only the boundary conditions governing the test must be modeled explicitly by means of time functions. These are the sea level fluctuation for the tidal response and the prescribed flow rates for the pumping tests. Boundary conditions are summarized in Table 3. Likewise, areal recharge does not need to be calculated (in fact, recharge in the entire study area is negligible).

| **Boundary** | **Type** | **Problem 1. Tidal Response** | **Problem 2. PT at OB-6** | **Problem 3. PT at OB-15** | **Problem 4. PT at OB-16** |
|---|---|---|---|---|---|
| East | Prescribed flow | Q=0 | Q=0 | Q=0 | Q=0 |
| West | Prescribed flow | Q=0 | Q=0 | Q=0 | Q=0 |
| South | Prescribed flow | Q=0 | Q=0 | Q=0 | Q=0 |
| North, seashore | Prescribed head | $\Delta h(t) = \Delta H_{sea}(t)$ | $\Delta h = 0$ | $\Delta h = 0$ | $\Delta h = 0$ |
| OB-6 | Prescribed flow | --- | Q = 100 l/s | --- | --- |
| OB-15 | Prescribed flow | --- | --- | Q = 100 l/s | --- |
| OB-16 | Prescribed flow | --- | --- | --- | Q = 100 l/s |

**Table 3** Boundary conditions of the geostatistical model. PT denotes pumping (and recovery) test. $\Delta H_{sea}(t)$ is depicted by dots in Fig. 2.

## 3.8. Spatial variability of unknown fields

Transmissivity and storage coefficient fields are highly heterogeneous, as revealed by the geophysical campaign (Fig. 1) and the prior estimation of effective diffusivities (Table 2). We use a geostatistical model consisting of a set of hard measurements (Table 2) and a correlation (covariance) structure arising from a preliminary study. The covariance structure is represented by a single anisotropic exponential variogram, without nugget effect and sill of 0.011 ($\log_{10}(m^2/s)$). Ranges are 600 m along the direction of larger correlation (N4W, Fig. 1) and 290 m in the

orthogonal one. This covariance structure applies to both transmissivity and storage coefficient, as they are assumed to be highly correlated. The spatial variability of these parameters is characterized using the regularized pilot points method (Alcolea, *et al.*, 2006). To this end, an unstructured network of pilot points (69 for each unknown field) has been designed. These are clustered in the zone encompassing the existing wells, where the majority of the information comes from (Fig. 4).

## 3.9. Results

Results are evaluated in terms of fits to available head variation measurements and plausibility of the solutions. The latter is evaluated both qualitatively (by visual comparisons with the resistivity map in Fig. 1) and quantitatively. To this end, we compare the diffusivities estimated by the model with those obtained by TRM (Table 2). Four out of two hundred simulated transmissivity and diffusivity fields (and the corresponding fields obtained by conditional estimation) are displayed in Figs. 5 and 6, respectively. Note that the unknowns are transmissivity and storage coefficient. However, observation of diffusivity ($D = T/S$) facilitates interpretation. In the 200 conditional simulations and the 'single best' conditional estimation, the monitored wells lay on a zone of medium diffusivity, connected to the sea and embedded between two parallel channels of high diffusivity. These present a clear orientation toward North and are also well connected to the sea. All characterizations reveal the presence of a low diffusivity zone close to the seashore. This can be explained by the deposition of fine, less permeable, materials along the coast line in the study area, where the marine currents are normally weak. Possibly, a Cauchy type boundary condition (i.e., leakage) at the seashore would have modeled this effect better. Yet, the fine discretization used close to the seashore (Fig. 4) helps to alleviate this problem.

Fits to head variation data are displayed in Fig. 7. They are all satisfactory, even for the simulation yielding the worst match to measured head variations. Very similar fits are obtained by conditional estimation. In spite of the smoothness of the estimated transmissivity and storage coefficient fields, conditional estimation captures the large scale patterns of heterogeneity, which are known to control groundwater flow (Alcolea, *et al.*, 2008). However, seeking a 'single best' characterization by conditional estimation is not a good option because it does not allow us to evaluate uncertainty. The quality of the fits to measured head variations is best observed in Fig. 8. There, we depict the Box and Whisker plots (boxplots hereinafter) of the averaged-in-time standardized residuals for all conditional

simulations and all observation wells and problems listed in Table 1. Normalization factors are the amplitude of the tide (2.62 m) and the maximum drawdown caused by pumping at wells OB6, OB15 and OB16 (1.22, 7.99 and 3.35 m, respectively). In addition, mean residuals obtained by conditional estimation are depicted. Three observations arise from Fig. 8. First, all boxplots are centered on a small value, regardless of the forcing term. Thus, the median of the residuals is small. This reflects the good quality of the fits for all the realizations. Second, mean residuals obtained by conditional estimation are, in general, slightly larger than those obtained by conditional simulation. However, these differences are not significant in any case. Third, the wings of the boxplots are short. This manifests the striking similarity between the simulated fields displayed in Figs. 5 and 6 (i.e., similar fits are obtained by different realizations). None of the simulations deviate significantly from the field obtained by conditional estimation. This convergence toward similar fields is even more striking when considering that the 200 initial simulations, conditioned only to T and S data (not depicted here) were all very different. After conditioning to head variations, they all became very similar. We argue that the information contained in the head variation measurements is sufficient to reduce uncertainty.

From a qualitative point of view, all transmissivity (and diffusivity) fields resemble vaguely the directional features of connectivity as observed in the resistivity map (Fig. 1). This comparison also reveals the presence of highly conductive bodies well connected to the sea in the resistivity map and all simulations. We further tested the physical plausibility of the simulated fields by comparing the values of effective diffusivity obtained by the TRM (see section ‘‘Analysis of the tidal response”) with those obtained by the stochastic model (Table 2). It is worth mentioning that the former data set was not used as conditioning data for the geostatistical inversion. The fact that both the TRM and the geostatistical model yield similar values of effective diffusivity is a good indication of the physical plausibility of the simulations.

From the above we conclude that the available data sets and the techniques used for aquifer characterization are available to identify the main patterns of heterogeneity. The surprising finding was the limited degree of variability of the simulated fields. Yet, there is still room for uncertainty due to errors in the conceptual model that were not accounted for in this study (e.g., the choice of correlation structure defining the geostatistical model, the boundary conditions and the position of the seashore, etc.).

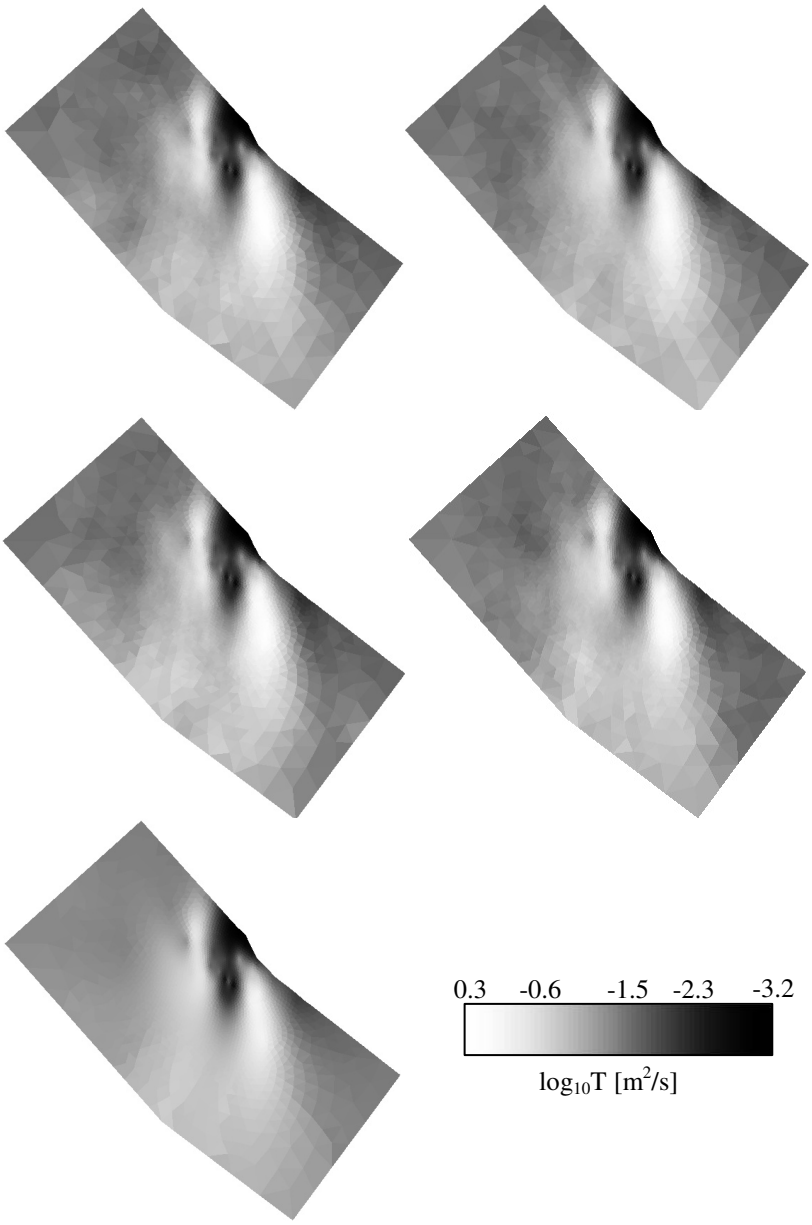



**Fig 5** Four out of two hundred conditionally simulated (above) and estimated (below) log-transmissivity fields.

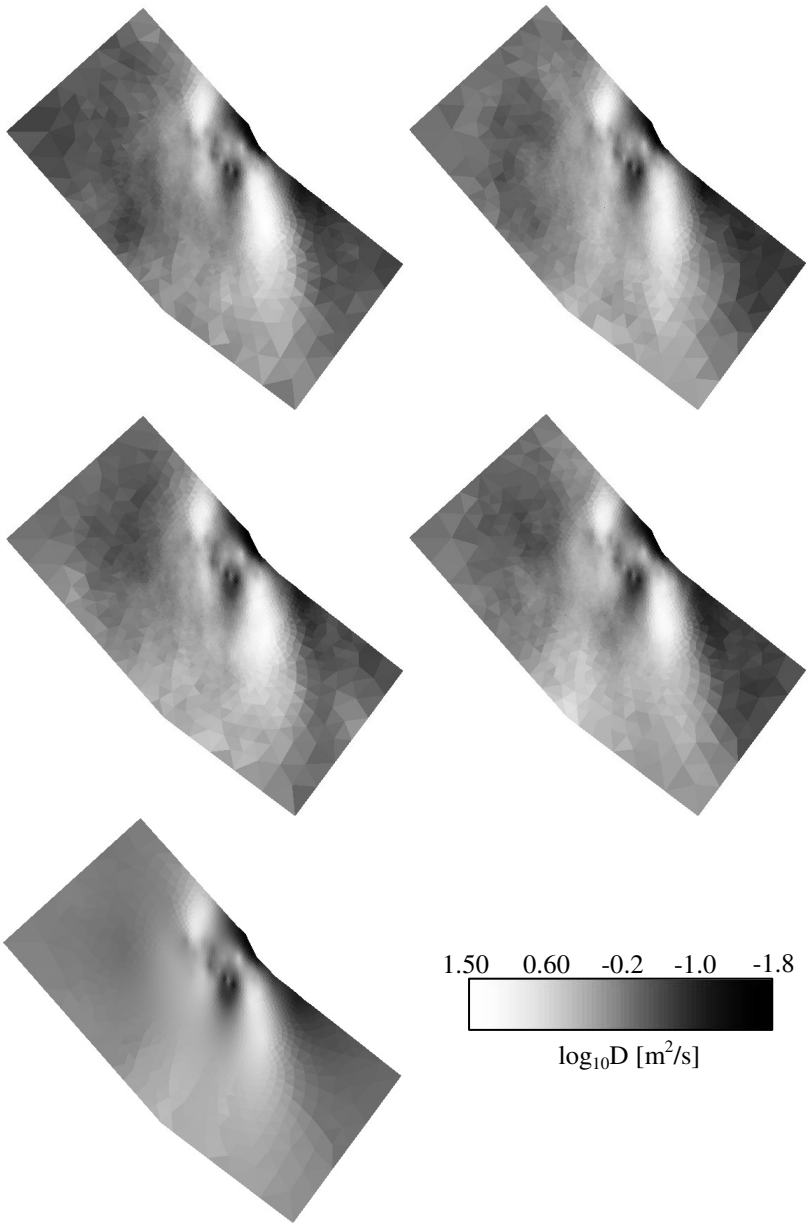



**Fig 6** Four out of two hundred simulated (above) and estimated (below) log-diffusivity fields, calculated as D=T/S.

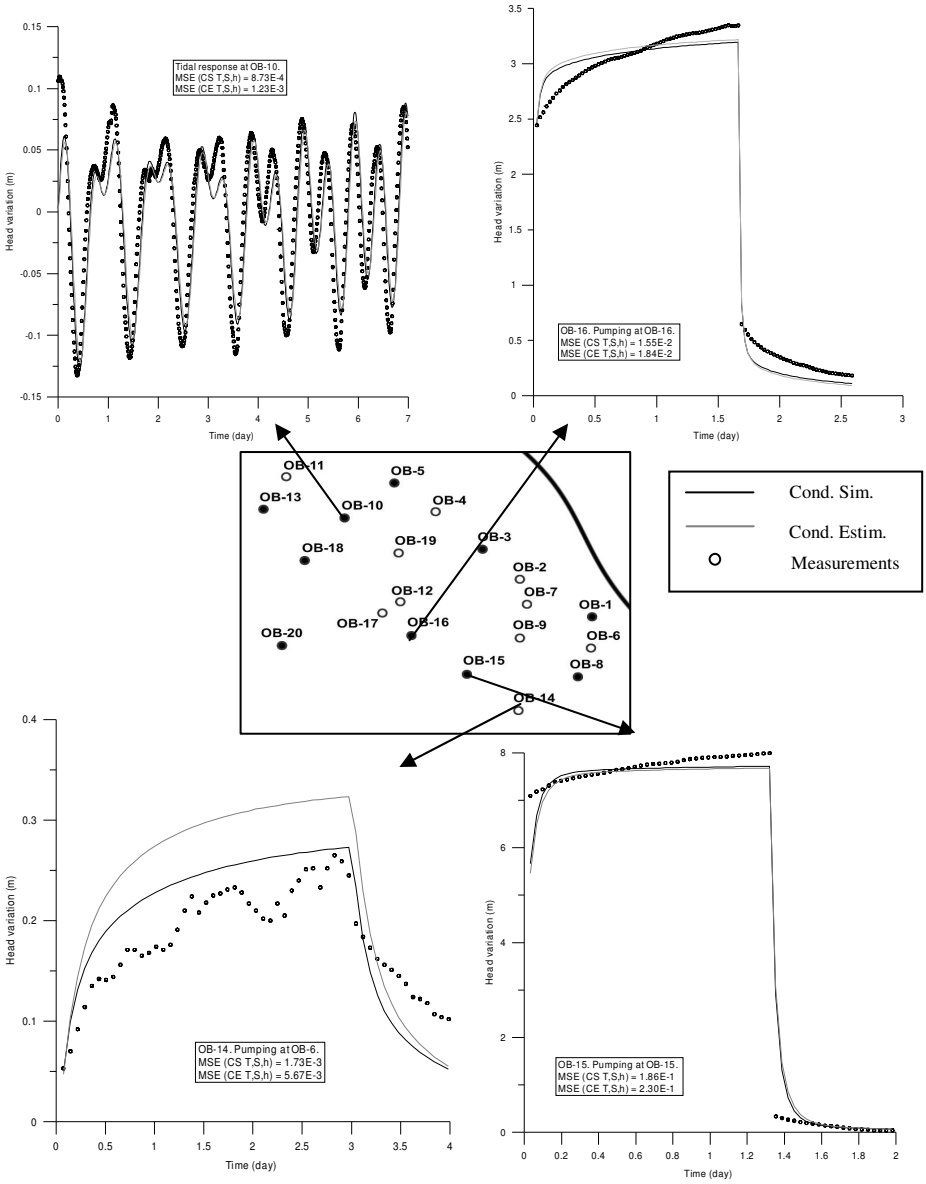



**Fig 7** Calculated and measured head variations at selected points in response to tidal and pumping effects. In the insets, MSE denotes Mean Square Error (Mean square difference between calculated and measured head variation). CS and CE denote conditional simulation and conditional estimation, respectively.

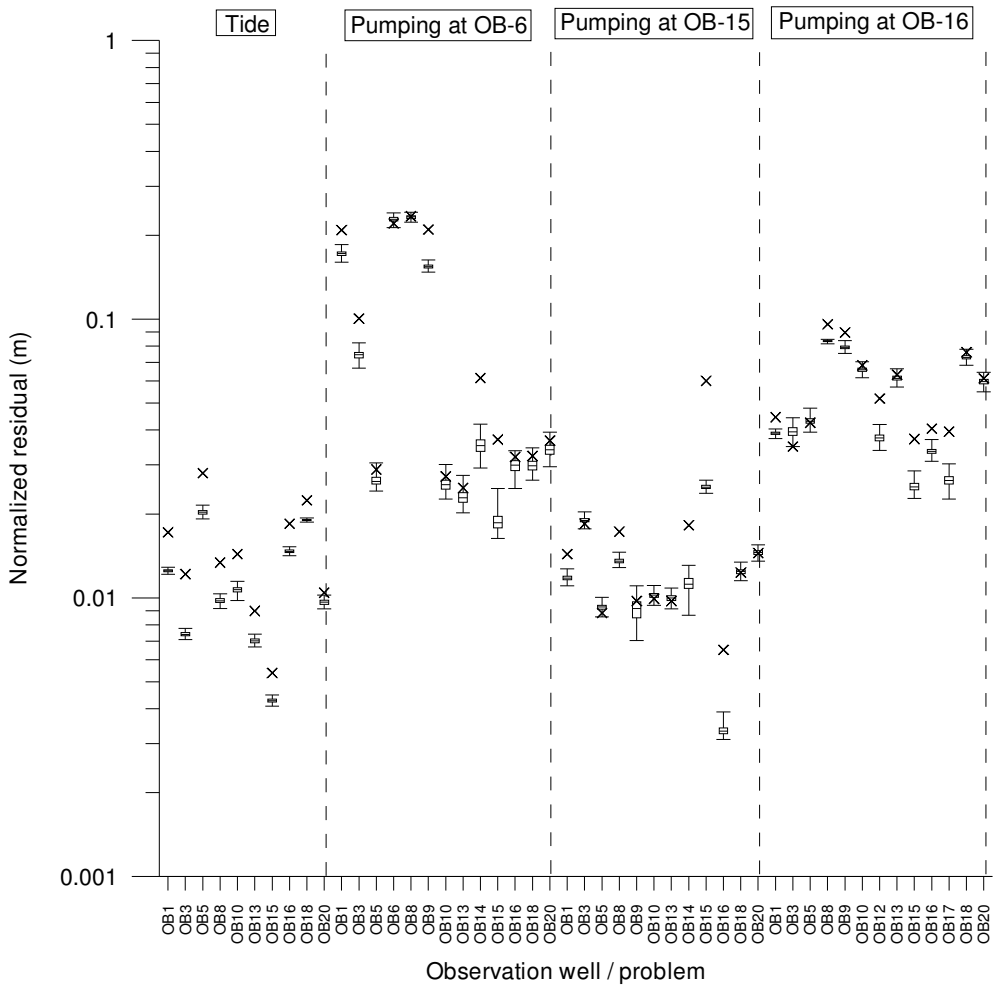



**Fig 8** Box and Whisker plots of standardized residuals (averaged in time). Normalization factors were the amplitude of the tide (2.62 m) and the maximum drawdown at the pumping well (1.22, 7.99 and 3.35 m for pumping at wells OB-6, OB-15 and OB-16, respectively). Crosses depict mean standardized residuals obtained by conditional estimation.

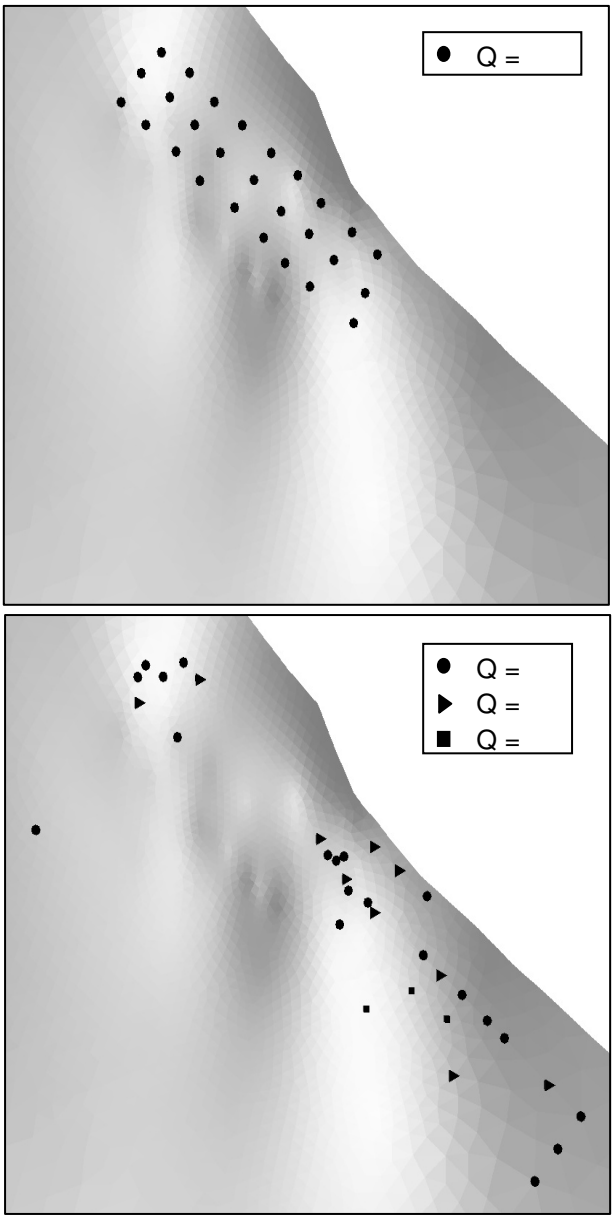



**Fig 9** Synthetic (above) and optimum (below) pumping networks. On the background, a shaded relief of the average diffusivity field.

## 4. Optimum pumping network

The 200 equally likely hydraulic characterizations are used to find a unique optimally-robust design of the pumping network, defined by the number of wells, their location and the corresponding discharge rates. In this optimization framework, a robust design (Watkins and McKinney, 1997) is defined such that, first, it satisfies the design constraints for all the hydraulic characterizations generated at the previous step. Second, it minimizes the total expected costs of set up, operation and maintenance of the solution (mean over all the simulations). More precisely, constraints of the design are:

- A target discharge of 9000 $m^3$/h. In fact, the solution is slightly over-designed, in order to ensure the groundwater supply to the desalination plant, so as to warrant the required production of 3346 m3/h of freshwater. This also accounts for potential stops at some wells for maintenance or repairing. Finally, the applied target discharge is 10,000 $m^3$/h.
- The solution must be technically feasible. Only three types of pump capable of handling highly aggressive brackish water with pumping rates of 360, 252 and 108 $m^3$/h are available.

The objective is then to minimize the impacts of pumping and the costs of drilling, maintenance and exploitation. This is partly achieved by minimizing drawdowns, which reduces impacts to the aquifer, electrical costs, risks of well collapse, etc.

### 4.1. Problem formulation and genetic algorithm

Under the aforementioned constraints, we solve a discrete (only three pumping rates can be considered), non linear (i.e., costs are not a linear function of pumping rates) optimization problem in a stochastic framework. Two techniques are applied. On the one hand, we used genetic algorithms to solve the aforementioned problem (Goldberg, 1989; Siegfried and Kinzelbach, 2006). On the other hand, a stacking approach (Chan, 1993; , 1994; Feyen and Gorelick, 2004; Morgan*, et al.*, 1993; Wagner and Gorelick, 1987) accounts for the inherent stochastic uncertainty. In order to keep calculation times reasonable, 126 mesh nodes have been selected as potential locations for the wells. Nodes at existing wells (Fig. 1) are included in the set of potential well locations. These are located in the highly diffusive bodies (Fig.

6) and within the lot owned by the desalination company. The 126 unknowns are the type of pump at each potential well, ranging from 0 (no pump) to 3 (maximum flow rate of 360 $m^3$/h). These are arranged as a vector of integers **x**, termed individual.

By virtue of the superposition principle, one can compute the drawdowns associated to a given individual **x** with the aid of a response matrix **A** (i.e., the element $a_{ij}$ being the drawdown at well '*i*' in response to a unitary pumping at potential well '*j*'). The components of **A** are calculated by running the model for a given transmissivity field successively considering all potential locations for the pumping wells. A steady state regime with pumping was used to consider the worst case situation. Only 100 out of 200 transmissivity fields were considered for the optimization network, as increasing the stack size adds more constraints, which makes it more tedious to find a robust solution (Feyen and Gorelick, 2004). In fact, the stack size was reduced due to CPU considerations (an optimization run using 100 transmissivity fields already takes 24 h in a high performance computer). A stacked matrix $\mathbf{A}^{\text{stack}}$ (Feyen and Gorelick, 2004) accounts for the response matrices corresponding to the 100 transmissivity fields. Drawdowns can then be expressed as

$$\mathbf{s} = \mathbf{A}^{stack}\mathbf{Q}(\mathbf{x}) \tag{3}$$

where **s** is a vector containing the drawdowns at all potential wells and for all stochastic simulations due to the pumping configuration **x** and **Q** is a vector containing the flow rates associated to the categories defined by the individual **x**. The use of $\mathbf{A}^{\text{stack}}$ warrants that the constraints are met for all considered transmissivity fields. Given an individual **x** and its associated drawdowns, the cost function over all wells and all stochastic simulations can be defined as

$$C = \sum_{j=1}^{Nsimu} \sum_{i=1}^{Nwells} (D_i + P_i + E_{i,j}) \tag{4}$$

where 'Nwells' and 'Nsimu' denote the number of potential wells and stochastic simulations, respectively. $D_i$, $P_i$ and $E_{i,j}$ are the drilling cost, the cost of the pump and its maintenance and the electrical costs at well '*i*' and simulation '*j*', respectively. Note that only the electrical costs depend on the simulation. Calculating D and P is straightforward, as they depend only on the category of the pump and the position (already existing wells have a drilling cost equal to zero). Electrical costs can be expressed in terms of potential energy of groundwater $E_p$:

$$Ep_i = mgh = (Q_i \rho t) g s_i \quad [joule] \tag{5}$$

where $Ep_i$ is the potential energy at well '*i*', m is mass of water [kg], $g$ is gravity [m/s$^2$], h is the required height of elevation [m], $Q_i$ is the flow rate at well '*i*' [m$^3$/s], $\rho$ is density (1032 kg/m3), $t$ is time [s] and $s_i$ [m] is the drawdown at well '*i*'. Time is set to 22 years, the expected life of the desalination plant. Finally, grouping all constant terms, $E_i$ is expressed as

$$E_i = Ep_i[kWh]\beta[euro/kWh] = Q_i s_i C_{unit} \tag{6}$$

where $\beta$ denotes the unitary cost of electricity (0.0399 Euro/kWh) and $C_{unit}$ is 77169 Euro/(m$^4$/s). Fixed costs and other parameters of the genetic algorithm are summarized in Table 4.

| Value | | | |
|---|---|---|---|
| $C_{unit}$ | | 77169 euros / (m$^4$/s). Assuming 22 years of exploitation and an unitary cost of 0.0399 euros/kWh. | |
| Drilling cost (D) / Cost of pump (P) | Category | CD (euros / well) | CD + CP (euros / well) |
| | 100 l/s | 23650 | 224327 |
| | 70 l/s | 23650 | 190678 |
| | 30 l/s | 0 (existing wells) | 100000 |
| Number of potential wells | | 126 | |
| Population at each generation | | 30 | |
| Number of generations | | 250000 | |
| Maximum drawdown allowed | | 15 m | |
| Penalty factor for drawdowns | | 100 | |
| Target discharge | | 10000 m$^3$/h (security factor of 1.111) | |
| Penalty factor for discharge | | 1000 | |

**Table 4** Parameters required by the genetic algorithm utilized for the optimization of the pumping network.

Posed in this way, the problem of minimizing the electrical costs $E$ is equivalent to that of minimizing the drawdown at the pumping wells. Once the individual **x** has been designed, additional constraints due to target discharge and maximum allowed

drawdown are addressed by multiplying the cost function by a penalty factor when these criteria are not met. Actually, the maximum drawdown allowed (15 m in this case) was not attained for any simulation in the stack. The optimum individual **x** is determined by the GaMin Matlab toolbox (see a detailed description in Popov, 2005). After performing a sensitivity analysis to study the convergence and the reliability of the solution, the size of the population has been set to 30 individuals. We used a conventional scattered cross-over mechanism. The mutation range and the part of the population copied to the next generation (elite coefficient) have been set to 20% and 7%, respectively. The part of the population with largest objective functions is replaced by new individuals (7%). Finally, the number of generations was set to 250,000.

| | | Minimum | Maximum | Average | Standard deviation |
|---|---|---|---|---|---|
| **Cost function (million euros)** | **Optimum** | 8,75 | 9,23 | 8,98 | 0,09 |
| | **Synthetic** | 9,65 | 10,0 | 9,82 | 0,06 |
| **Drawdown at pumping wells (m)** | **Optimum** | 4.76 | 14.65 | 9.47 | 1.57 |
| | **Synthetic** | 7.71 | 37.04 | 17.08 | 7.28 |
| **Drawdown at all nodes (m)** | **Optimum** | 0.00 | 14.65 | 8.36 | 2.03 |
| | **Synthetic** | 0.00 | 37.04 | 10.89 | 4.43 |
| **Maximum drawdown at nodes inland (m)** | **Optimum** | 7.93 | 10.38 | 9.15 | 0.48 |
| | **Synthetic** | 8.85 | 10.87 | 9.87 | 0.37 |

**Table 5** Summary of statistics of the comparison between the synthetic and the optimum pumping networks.

## 4.2. Optimization results

The benefit of optimization is analyzed by comparing two pumping configurations. First, we test a tradition based, hand-delineated, pumping network arising from a preliminary study not accounting for optimization. In that solution, it was planned to locate 27 pumping wells along three lines parallel to the seashore, following current practice. Pumping rate is identical at all wells (i.e., target discharge of 10000 $m^3$/h divided by the number of wells). Second, we optimize the pumping configuration for 100 transmissivity fields. Results are evaluated in terms of cost of the solution and drawdowns at the pumping wells, at all nodes of the finite element mesh and at those defining the inland boundary. Fig. 9 displays the synthetic and optimum pumping configurations. Fig. 10 displays the cumulative distribution functions (cdf hereinafter) of costs and drawdowns. Statistics of those cdfs are summarized in Table 5. It is worth emphasizing that the optimization was performed using 100 transmissivity fields. However, the performance of the optimum and synthetic networks is evaluated for the complete stack of 200 realizations. The optimum transmissivity field obtained by conditional estimation was not considered for optimization as it does not allow us to evaluate the uncertainty of the suggested solution.

The optimum distribution of wells (and corresponding flow rates) is reasonable, as observed in Fig. 9. In general, largest pumping rates of 360 $m^3$/h correspond to wells along the two high diffusivity channels. This causes little drawdowns and a superior yield of the system. Flow rates of 252 and 108 $m^3$/h are assigned mainly to existing wells. We attribute this to the fact that drilling costs are zero at existing wells.

The effect of the optimization is best analyzed observing Fig. 10 and Table 5. First, the total cost of the system is reduced substantially (Fig. 10a). In average, the reduction is of about 10% of the total cost. The underlying uncertainty is very small, as measured by the standard deviation of the cost function. Second, it reduces significantly the drawdowns (at the pumping wells, at all nodes and at those defining the inland boundary; Fig. 10 b, c and d, respectively). Thus, the optimum pumping configuration minimizes the total cost and the inherent environmental hazards. This effect is best observed in Table 5. On average, the optimum configuration reduces the drawdown at pumping wells by approximately 8 m. This minimizes both the risk of the pump failure and of well collapse. This reduction is not dramatic in terms of generalized drawdowns or drawdowns at nodes defining inland boundary (average reduction of 2.5 m and 0.75 m, respectively). These magnitudes are more sensitive

to the amount of water being removed rather than to the location/configuration of the removal. A maximum drawdown of 10 m (30 m for the synthetic network) is barely achieved at a few transmissivity fields which were not considered for the optimization. This confirms that the optimum pumping network also accounts for minimum aquifer vulnerability.

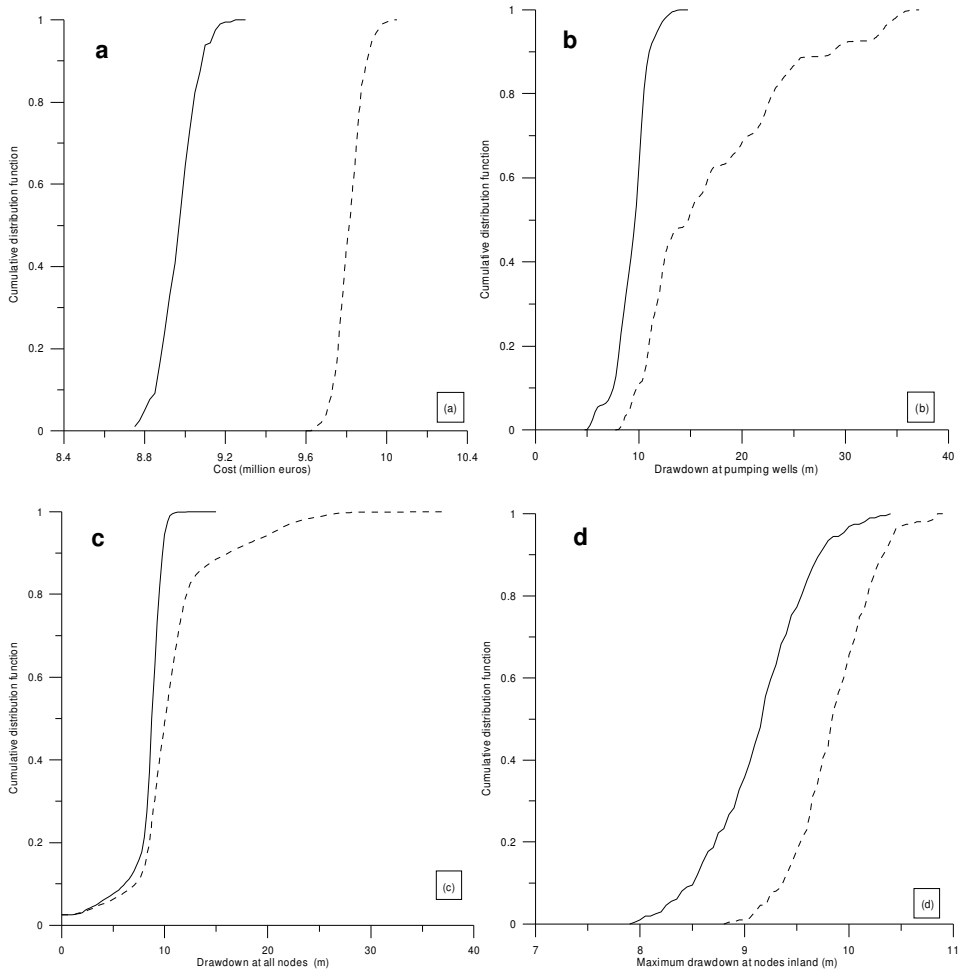



**Fig 10** Cumulative distribution functions of optimum (solid line) and synthetic (dashed line) pumping networks: (a) cost function, (b) drawdown at

pumping wells, (c) drawdowns at all nodes of the finite element mesh and (d) drawdown at nodes defining the inland boundary of the model.

## 5. Conclusions

This work summarizes the application of stochastic inverse modeling and optimization techniques to the management of a pumping system at a coastal aquifer in Oman. A stochastic characterization of hydraulic parameters from tidal fluctuation and pumping test data was used to design an optimum pumping network of brackish groundwater. This would allow an increase in the current production of a desalination plant, which will satisfy the growing demand of freshwater within the area.

The applied methodology consists of two main steps. First, transmissivity and storage coefficient fields are characterized from available data using a stochastic model. Spatial variability of these parameters is addressed by the regularized pilot points method. We obtain 200 equally likely simulations of the transmissivity and storage coefficient fields that are plausible (i.e., fit the diffusivities obtained by TRM and resemble the connectivity features of a resistivity map obtained by geophysics) and fit well the indirect head variation measurements.

Second, an optimum pumping network is designed with the help of a genetic algorithm. The aim is to obtain a reliable solution that minimizes the expected cost and that allocates the pumping wells in such a way that the drawdowns are small. This minimizes the electrical costs and the environmental side effects. Constraints of the design are the target discharge (10,000 $m^3$/h) and the technical feasibility of the solution (i.e., predefined pumping rates of 360, 252 and 108 $m^3$/h).

The performance of the optimum solution is compared to the one of a synthetic, tradition based, hand-delineated, pumping network. Results show that the use of the optimization technique leads to a reduction in operational costs of more than 10%. In addition, drawdowns are reduced dramatically. Certainly, this cannot be considered as a fair comparison, since different reduction factors could be obtained when comparing the optimum solution with other configurations, traditionally designed in a subjective way by water managers.

Much remains to be done. Uncertainties on the conceptual model were neglected in this work. Among them, the location of the contact between aquifer and sea may have a large impact on the calibration results. Geophysics and tracer tests can provide valuable information for the identification of the karstic conduits. These data sets could be considered as conditioning data for the stochastic model. Non

multi-Gaussian techniques, such as multiple point geostatistics could be useful to infer patterns of heterogeneity in a more realistic manner. Yet, this work demonstrates the value of combining stochastic inverse modeling to infer aquifer heterogeneity and optimization techniques to determine (objectively) the optimum pumping configuration. This is a promising methodology for designing pumping networks in highly heterogeneous aquifers while also minimizing the environmental impacts of desalination plants.

## 6. References

Abarca, E., Vazquez-Suñe, E., Carrera, J., Capino, B., Gamez, D., and Batlle, F. (2006), *Optimal design of measures to correct seawater intrusion*, Water Resources Research, 42, 9, W09415.

Ahlfeld, D., and Heidari, M. (1994), *Applications of optimal hydraulic control to groundwater systems*, Water Resour. Res., 120, 3, 350-365.

Ahlfeld, D., and Mulligan, A. (2000), *Optimal Management of Flow in Groundwater Systems*

Alcolea, A., Carrera, J., and Medina, A. (2006), *Pilot points method incorporating prior information for solving the groundwater flow inverse problem*, Advences in Water Resources, 2006, 29, 1678-1689.

Alcolea, A., Carrera, J., and Medina, A. (2008), *Regularized pilot points method for reproducing the effect of small scale variability. Application to simulations of contaminant transport*, Journal of hydrology, 355, 1-4, 76-90.

Alcolea, A., Castro, E., Barbieri, M., Carrera, J., and Bea, S. (2007), *Inverse modeling of coastal aquifers using tidal response and hydraulic tests*, Ground Water, 45, 6, 711-722.

Bugeda, G. (1990), *Utilización De Técnicas De Estimación De Error Y Generación Automática De Mallas En Procesos De Optimización Estructural*, Technical University of Catalonia, Barcelona.

Carrera, J., and Neuman, S. (1986a), *Estimation of aquifer parameters under transient and steady-state conditions, 2. Uniqueness, stability and solution algorithms.*, Water Resour. Res., 22, 2, 211-227.

Carrera, J., and Neuman, S. (1986b), *Estimation of aquifer parameters under transient and steady-state conditions, 3. Applications.*, Water Resour. Res., 22, 2, 228-242.

Chan, N. (1993), *Robustness if the multiple realization method for stochastic hydraulic aquifer management*, Water Resour. Res., 29, 9, 3159-3167.

Chan, N. (1994), *Partial infeasibility method for change-constrained aquifer management*, J. Water Resour. Plan. Manage., 120, 1, 70-89.

Chen, C., and Jiao, J. (1999), *Numerical simulation of pumping tests in multilayer wells with non-Darcian flow in the well-bore*, Ground Water, 37, 3, 465-474.

Cheng, A., Halhal, D., Naji, A., and Ouazar, D. (2000), *Pumping optimization in saltwater-intruded coastal aquifers*, Water Resour. Res., 36, 8, 2155-2165.

Cornaton, F. (2006), *GroundWater, A 3-D Ground water flow and transport finite element simulator*, Univ. Neuchâtel.

Delyanns, E. (2003), *Historic background of desalination and renewable energies*, Solar Energy, 75, 357-366.

Dickie, P. (2007), Desalination: option for distraction for a thirsty world?, 53 pp, WWF Global Freshwater Programme.

Ferris, J. (1951), *Cyclic fluctuations of water level as basis for determining aquifer transmissibility*, in *International Association of Hydrological Sciences, Publication 33*, pp. 148-155, IAHS.

Feyen, L., and Gorelick, S. (2004), *Reliable groundwater management in hydroecologically sensitive areas*, Water Resour. Res., 40, W07408.

Goldberg, D. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.

Gorelick, S. (1983), *A review of distributed parameter groundwater modeling methods*, Water Resour. Res., 19, 2, 305-319.

GWI (2005), *19th IDA Worldwide Desalting Plant Inventory*, Media Analytics Limited (Global Water Intelligence).

GWI (2006), *Desalination Markets 2007: A Global Industry Forecast*, Media Analytics Limited (Global Water Intelligence).

Hendricks-Franssen, H.-J., Gomez-Hernandez, J., Sahuqillo, A., and Capilla, J. (1999), *Joint simulation of transmissivity and storativity fields conditional to hydraulic head data*, Advances in Water Resources, 23, 1-13.

Hug, R. (2005), *Hydrogeologische Untersuchungen im Abstrom der Sondermülldeponie Kölliken*, Bull. angew. Geol., 10, 1, 65.

Hvorslev, M. (1951), *Time-lag and soil permeability in ground-water observations*, US Army Corps of Engineers.

Li, W., Englert, A., Cirpka, O., Vanderborght, J., and Vereecken, H. (2007), *Two-dimensional characterization of hydraulic heterogeneity by multiple pumping tests*, Water Resour. Res., 43, 4, W04433.

Mantoglou, A. (2003), *Pumping management of coastal aquifers using analytical models of saltwater intrusion*, Water Resour. Res., 39, 12, 185-201.

Medina, A., Alcolea, A., Carrera, J., and Castro, L. (2000), *Modelos de flujo y transporte en la geosfera: Codigo Transin IV. [Flow and transport modelling in the geosphere: the code TRANSIN IV]*, in *IV Jornadas de Investigacion y Desarrollo Tecnologico de Gestion de Residuos Radiactivos de ENRESA*, pp. 195-200, ENRESA.

Medina, A., and Carrera, J. (2003), *Geostatistical inversion of coupled problems: dealing with computational burden and different types of data*, Journal of hydrology, 281, 251-264.

Meier, P., Carrera, J., and Sanchez-Vila, X. (1998), *An evaluation of Jacob's method work for the interpretation of pumping tests in heterogeneous formations*, Water Resour. Res., 34, 5, 1011-1025.

Morgan, D., Eheart, J., and Vallocchi, A. (1993), *Aquifer remediation design under uncertainty using a new chance constrained programming technique*, Water Resour. Res., 29, 3, 551-561.

Popov, A. (2005), *Genetic Algorithms for optimization, User Manual*, Hamburg.

Popov, A., and Filipova, K. (2004), *Genetic Algorithms - synthesis of finite state machines*, paper presented at 27th International Spring seminar on electronics technology, Herlany (Slovak Republic).

Renard, P. (2007a), *HYTOOL c1.91. User's guide*

Renard, P. (2007b), *Stochastic hydrogeology: what professionals really need?*, Groundwater, Manuscript submitted.

Rotting, T., Carrera, J., Bolzicco, J., and Salvany, J. (2006), *Stream-stage response tests and their joint interpretation with pumping tests*, Ground Water, 44, 3, 371-385.

Siegfried, T., and Kinzelbach, W. (2006), *A multiobjective distre stochastic optimization approach to shared aquifer management: methodology and application*, Water Resour. Res., 42, W02402.

Tarantola, A. (2005), *Inverse Problem Theory and Methods for Parameter estimation*, Society for Industrial and Applied Mathematics.

Trefry, M., and Johnston, C. (1998), *Pumping test analysis for a tidally forced aquifer*, Ground Water, 36, 3, 427-433.

Wagner, B. (1995), *Recent advances in simulation optimization groundwater management modeling*, Rev. Geophys, 33, 1021-1028.

Wagner, B., and Gorelick, S. (1987), *Optimal groundwater quality management under parameter uncertainty*, Water Resour. Res., 23, 7, 1162-1174.

Watkins, D., and McKinney, D. (1997), *Finding robust solutions to water resources problems*, J. Water Resour. Plan. Manage., 123, 1, 49-58.

Weiss, R., and Smith, L. (1998), *Parameter space methods in joint parameter estimation for groundwater flow models*, Water Resour. Res., 34, 4, 647-661.

# Appendix C

## Improving the performance of the Direct Sampling algorithm

*"From the moment we enter this life, we are in the flow of it [the time]. We measure it and we mark it, but we cannot defy it. We cannot even speed it up or slow it down. Or can we?"*

*Edward Norton in The Illusionist (2006)*

# 1 Introduction

This appendix is aimed at identifying the relevance and the sensitivity of the parameters governing the performance of the Direct Sampling (DS) algorithm presented in chapter 2, and to discuss some improvements that can be brought to the method.

The first part provides a detailed description of the algorithm as implemented in the C++ code used for the examples displayed in the body of the thesis. It is followed by a sensitivity analysis of the algorithm parameters. Then comes a discussion on where the CPU bottlenecks are and propositions for accelerating the algorithm. Finally, an updated description of the algorithm is presented, including the proposed changes.

# 2 Detailed description of the Direct Sampling algorithm

**Define:**

| | |
|---|---|
| $z$ | Simulated variable. |
| $n$ | Maximum nb. of nodes in a given data event. |
| $t$ | Distance threshold. |
| $\delta$ | Distance weighting function. |
| $f$ | maximum fraction of TI to scan |
| $\mathbf{x}$ | Vector describing the position of a node in the SG (simulation grid). |
| $\mathbf{y}$ | Vector describing the position of a node in the Training Image (TI) grid. |
| $\mathbf{L} = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\} = \{\mathbf{x}_1 - \mathbf{x}, .., \mathbf{x}_n - \mathbf{x}\}$ | Lag vectors defining the neighborhood of $\mathbf{x}$. |
| $\mathbf{N}(\mathbf{x}, \mathbf{L}) = \{\mathbf{x} + \mathbf{h}_1, .., \mathbf{x} + \mathbf{h}_n\}$ | Data event for a node $\mathbf{x}$. |
| $d\{\ \mathbf{d}_n(\mathbf{x},\mathbf{L}), \mathbf{d}_n\ (\mathbf{y},\mathbf{L})\ \}$ | Distance between the data event found in the SG and the one found in the TI. |

**Algorithm steps:**

1. Attribute conditioning data to SG.
2. Define a path in the SG (random or unilateral) that does not include the conditioning data.
3. Loop on each node **x** of the path:
   a) Find the neighborhood of **x** made of $n$ nodes, defined by a set of lag vectors **L**.
   b) If no neighbor is found for **x**, randomly take a node **y** in the TI and assign its value $z(\mathbf{y})$ zo $z(\mathbf{x})$.
   c) Define the data event $\mathbf{N}(\mathbf{x},\mathbf{L})$.
   d) Define the window in the TI where $\mathbf{N}(\mathbf{x},\mathbf{L})$ can fit.
   e) Randomly draw a location **y** in the scan window for an initial scanning point.
   f) Define the number of scanning attempts in the TI $s$.
   g) Loop $s$ times:
      i) Define **y** as the next node in the scan window. If the end of the scan window is reached, restart at the first node of the scan window.
      ii) Find the data event $\mathbf{N}(\mathbf{y},\mathbf{L})$.
      iii) Compute the distance $d\{ \mathbf{d}_n(\mathbf{x},\mathbf{L}), \mathbf{d}_n (\mathbf{y},\mathbf{L}) \}$ by looping on each of the $n$ nodes in $\mathbf{N}(\mathbf{y},\mathbf{L})$ and by using an appropriate measure of distance.
      iv) If $d\{ \mathbf{d}_n(\mathbf{x},\mathbf{L}), \mathbf{d}_n (\mathbf{y},\mathbf{L}) \}$ is the lowest distance for this simulated node **x**, keep $\mathbf{y}_{best}$ in memory.
      v) If $d\{ \mathbf{d}_n(\mathbf{x},\mathbf{L}), \mathbf{d}_n (\mathbf{y},\mathbf{L}) \}$ is under a certain acceptance threshold $t$, the node is accepted and its value $z(\mathbf{y})$ is attributed to $z(\mathbf{x})$. Otherwise, go to i.
      vi) If the number of iterations of the loop i-iv has exceeded $s$, the node $\mathbf{y}_{best}$ with the lowest distance is accepted and its value $z(\mathbf{y}_{best})$ is given to $z(\mathbf{x})$.
   h) End loop
4. End loop

## 3 Sensitivity analysis on the parameters governing Direct Sampling

Sensitivity of the parameters is evaluated using two criteria: 1) the quality of the resulting simulations and 2) the performance in terms of CPU time. The sensitivities are tested in a very crude way by adjusting parameters one at a time. Evaluating the quality of a simulation is not straightforward. Here, the evaluation is qualitative: we look at the reproduction of specific spatial features such as the shape and connectedness of certain geobodies (channels).

### 3.1 Training Image

The training image used for all the tests presented in this appendix is borrowed from (Zhang*, et al.*, 2008). It consists of a fluvial system with a categorical variable having 3 possible state codes: sand, silt and shale. Its size is 335 by 320 nodes. It is displayed in Figure 1.

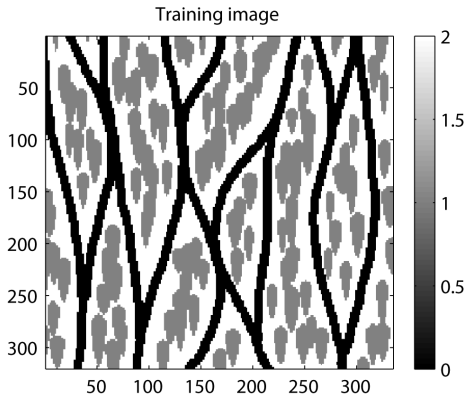



**Figure 1** Training image used for performance tests.

Sensitivity analysis is performed on the parameters *n, t, f and δ.*
The Fixed parameters are set as follows:

- Size of SG: 250 by 250 nodes.
- One simulation performed for each parameters set.
- No parallelization (only 1 CPU used).
- Random path.
- No syn-processing.
- Unconditional simulations (i.e. no hard data).
- Univariate simulations.
- Measure of distance: fraction of non-matching nodes in the data event (for categorical variable):

$$d\{\mathbf{d}_n(\mathbf{x}),\mathbf{d}_n(\mathbf{y})\} = \frac{\sum_{i=1}^{n} a_i \left\|\mathbf{h}_i\right\|^{-\delta}}{\sum_{i=1}^{n} \left\|\mathbf{h}_i\right\|^{-\delta}} \in [0,1], \quad where\ a_i = \begin{cases} 0 & if\ Z(\mathbf{x}_i) = Z(\mathbf{y}_i) \\ 1 & if\ Z(\mathbf{x}_i) \neq Z(\mathbf{y}_i) \end{cases} \qquad (1)$$

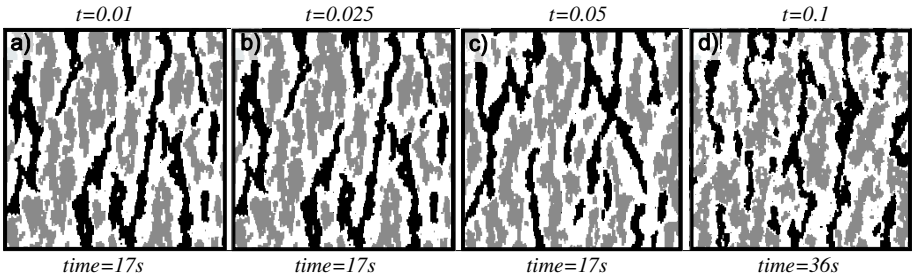
As a single simulation is performed for each parameters set, the calculation times and the resulting images are only a rough indication. Because the CPU time can vary for a given parameters set (depending on the random seed used), a large number of simulations would be necessary to obtain average times. Nevertheless, we believe that one simulation is enough to grasp the general trends driving the CPU times, to identify bottlenecks and to give clues on how to tackle them.

### 3.2 Sensitivity analysis on neighborhood (*n*) and threshold (*t*)

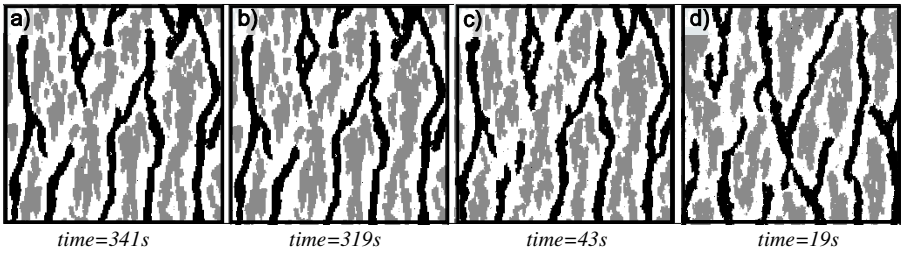
Figure 2 shows that both *n* and *t* are very important factors controlling the quality of simulations and the CPU time. Setting high *n* and low *t* yields the best simulations, but indeed this comes at the price of CPU time, because it involves scanning large portions of the TI to find a matching data event.

| Fixed parameters: | Varying parameters: |
|---|---|
| • $f = 0.5$ | • $n$ ranges from 10 to 40 |
| • $\delta = 0.5$ | • $t$ ranges from 0.01 to 0.1. |

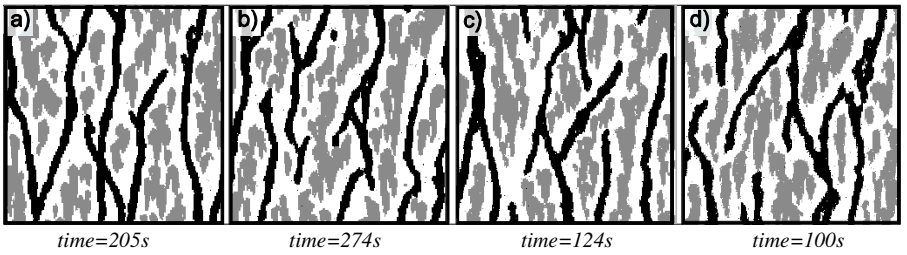
*n=10*




*n=20*




*n=30*




*n=40*

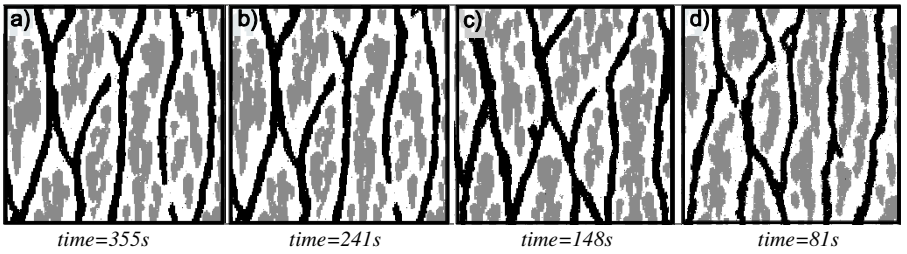


**Figure 2** Sensitivity analysis to parameters *n* and *t*

Setting *n* to low values results in discontinuity of geobodies and irregular shapes. Indeed, the complexity of the structures is not captured by small data events.

Setting *t* to a high value (e.g. 0.1) means to accept mismatching data events and to add noise. It is clearly visible on certain simulations, presenting noise and artifacts.

### 3.3 Sensitivity analysis on maximum fraction of TI to scan (*f*)

| Fixed parameters: | Varying parameters: |
|---|---|
| • $n = 30$ | • *f ranging from* 0.1 *to* 1 |
| • $t = 0.025$ | |
| • $\delta = 0.5$ | |

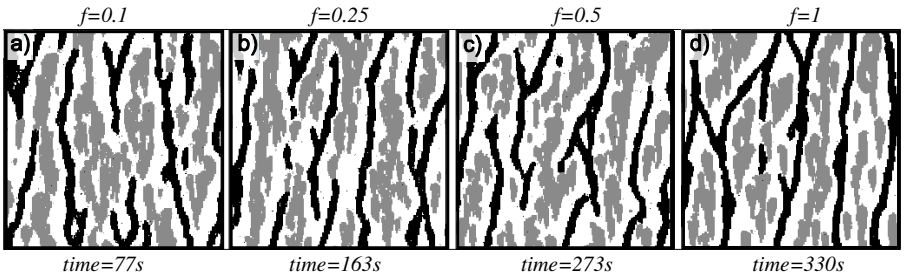



**Figure 3** Sensitivity analysis to parameter *f*

According to Figure 3, even though decreasing *f* dramatically reduces simulation times, it only mildly degrades the simulations.

Provided that the TI is stationary, repetitive and large enough, a portion of it is statically identical to another portion. Therefore, limiting the scan only causes errors in a few cases, when extremely rare data events are searched for.

### 3.4 Sensitivity analysis on distance weighting function ( $\delta$ )

| Fixed parameters: | Varying parameters: |
|---|---|
| • $n = 30$ | • $\delta$ *ranging from -10 to 10* |
| • $t = 0.025$ | |
| • $f = 0.5$ | |

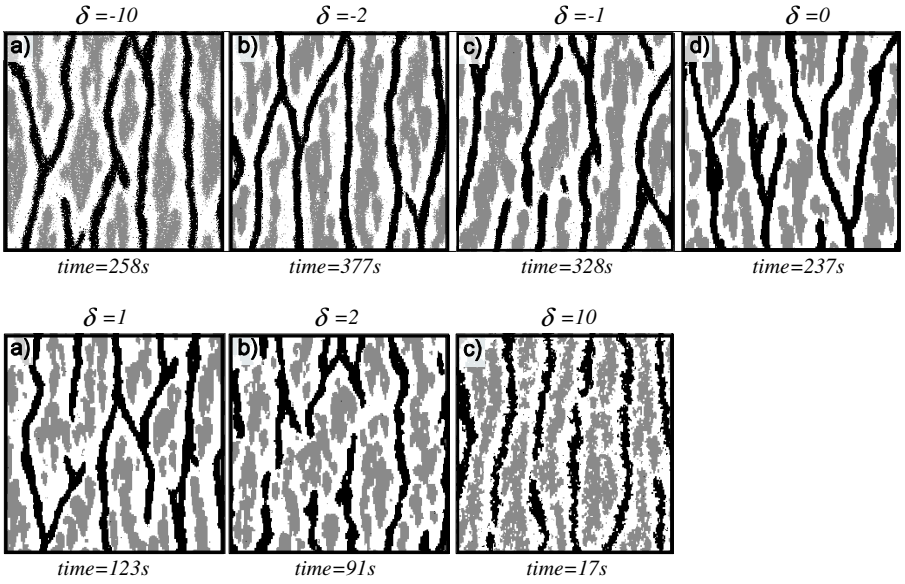



**Figure 4** Sensitivity analysis on parameter $\delta$

The parameter $\delta$ is controlling the relative importance of the template nodes as a function of their geometric distance to the central node (the node to simulate).

If $\delta$ is too large, only the central nodes of the data event are considered and it is equivalent to having a small $n$. Indeed, giving very little importance to distant nodes is equivalent to having smaller neighborhoods. Conversely, if $\delta$ is set to a negative value, the distant neighbors have more weight than the ones close to the central node. The large scale structures are then well reproduced, but the small scale is ignored, resulting in a noisy image.

More work could be dedicated to understanding how to address the issues of weighting the data event nodes. For example, kriging weights could be used. The advantage of this implementation is that a single parameter controls the weights, for any given data event.

## 4 Propositions for increasing performance

Various tests as well as code profiling showed that most of the simulation time is spent at step 3.g.iii, when computing the distance between two data events. The

consequence is that the larger the number of scanned TI nodes, the more time is needed.

If a large part of the TI has to be scanned, the computation time needed could make the simulation impractically long. The bet of DS is that a satisfying data event can be found quickly. This is illustrated in Figure 5. Figure 5a shows one simulation performed with the parameters indicated on top of the figure. Figure 5b displays the distance associated to each simulated node. This distance is higher at locations where complex structures appear in the image, such as for example on the sides of the channels. Similar features appear on figure 5c, displaying the number of scan attempts that have been performed for each simulated node (number of evaluations of the loop i-vi). Indeed, these two maps are related because when a data event is difficult to find in the TI and needs a lot of scan attempts, it is more likely that it will be only partially matched, incurring a higher distance.

Figure 5c shows the distribution of the number of scan attempts. In this case, DS seems to win its bet because most of the nodes have been simulated by scanning a very small portion of the TI. This is confirmed by the overall simulation time of 20s. The position of the gravity center of the histogram is proportional to the time consumed for the simulation. The main factors that may shift this gravity center to the right are high thresholds and large neighborhoods.

Several ways of reducing the amount of scan have been proposed. The first one is to interrupt it, by accepting non-optimal data events (setting a high threshold), but it degrades the quality of the results. Reducing the number of neighbors, by incurring less iterations of the loop i-vi, has a similar effect. It is also possible to refuse very infrequent data events by limiting the scan to a certain fraction of the TI.

Favoring the general coherence of all simulated nodes also fastens the simulation as it generates less data events that are incompatible with the training image, which involve scanning the entire scan window. Figure 2 shows an example of such situation where the simulation with $t = 0.1$ and $n = 10$ takes more time than the one with $t = 0.01$ and $n = 10$. This also explains why, in certain cases, syn-processing can speed-up the execution times while improving the simulation. (However, most of the time, syn-processing is time consuming because of its recursive nature).

| Parameters: | |
|---|---|
| • $n = 30$ | • $\delta = 1.0$ |
| • $t = 0.075$ | • $f = 1.0$ |

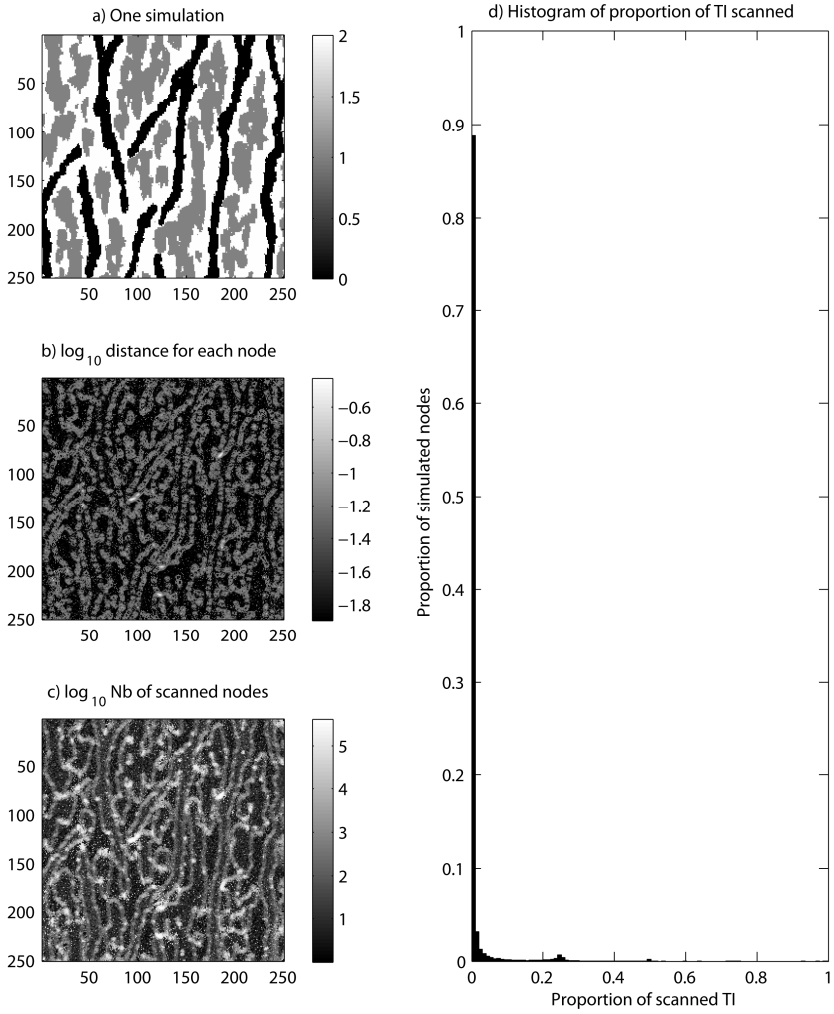



**Figure 5** a) Realization. b) Map representing the final distance associated to each simulated node. c) Map representing the number of scan attempts in the TI that have been necessary to simulate each node of the SG. d) Histogram of the number of scan attempts.

## 4.1 Scan interruption

The DS algorithm basically consists in 3 imbricated loops:

```
Loop all nodes of the simulation grid
    Loop all nodes of the training image grid
        Loop all nodes of the data event
                Compute distance
        End
    End
End
```

Even if the computation of the distance for one node of the data event is a small operation that can be accomplished in a few clock cycles, it is clear that using large data events is prohibitively time-consuming. Therefore, we propose to prematurely interrupt the loop on the $n$ nodes of the data event if it is obvious that a match will not be found. For example, if the threshold is set at 0.1 and the data event is made of 40 nodes, as soon as 4 non-matching nodes are found for one candidate data event, the entire data event can be considered non-matching and the scan can proceed to the next node. The interrupted loop is in the part 3.g.iii of the algorithm.

We introduce here a new parameter $i_s$, that defines the proportion of non-matching nodes where a scanned node can be considered as prematurely non-matching.

Tests presented in Figure 6 show that the simulation times are significantly reduced with no change in the resulting images. The explanation for the absence of degradation in the simulations quality is that the nodes whose scan was interrupted would not have been retained anyway with a complete scan. The only difference between the simulations shown in figure 6 is the CPU time needed. Scan interruption is therefore an efficient way to accelerate the simulation without changing its quality.

Nevertheless, when $i_s$ is too small (for example smaller than the threshold), it can happen that sometimes no TI node is retained at all during the entire scan. In the present implementation, when this happens the TI is re-scanned setting temporarily $i_s$ to 1.

| Fixed parameters: | Varying parameters: |
|---|---|
| • $n = 30$<br>• $t = 0.025$<br>• $\delta = 0.5$<br>• $f = 0.5$ | • $i_s$ *ranging from* 0.1 *to* 1 |

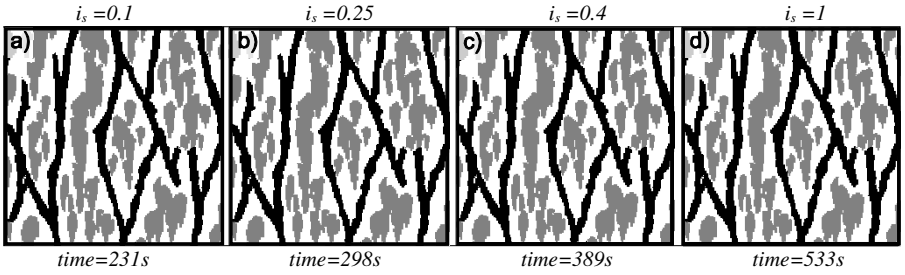



**Figure 6** Sensitivity analysis on parameter $i_s$

### 4.2 Post-processing

Both the discussion on the algorithm and the examples showed that the threshold value has a strong influence on the simulation time. High threshold values generate noisy simulations, but preserves quite well the large structures.

Therefore, a 2-step procedure is proposed. It consists in generating first a simulation with a high threshold, and then performing an additional step which consists in re-simulating all nodes, but this time with entirely informed neighborhoods. This post-processing operation can be carried on $p$ times, resulting in several post-processing passes. The idea is that the large structures can be simulated rapidly with a high threshold, at the price of creating noise. The noise is removed in the second step, where $n$ and $f$ can be divided by a certain factor $p_f$ to save CPU time in the additional passes. Figure 7 illustrates the quality improvement resulting of post-processing.

| Fixed parameters: | Varying parameters: |
|---|---|
| • $n = 30$<br>• $t = 0.1$<br>• $\delta = 0.5$<br>• $f = 0.5$<br>• $i_s$=*0.25* | • $p = 0$ *or* $1$ *post-processing pass with a factor* $p_f = 3$ |

*No post-processing* *One post-processing pass*

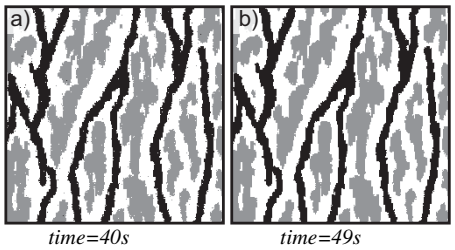



**Figure 7** Effect of post-processing on performance and quality

### 4.3 Parameters reduction

Parameters reduction is an *ad hoc* way of speeding up the simulations. The idea is that if the large and complex structures are difficult to determine at the beginning of the simulation, the small structures can be handled easier with smaller data events, higher threshold, etc. At a certain point in the simulation progress $P_r$ (determined as a percentage of simulated nodes), various parameters are reduced by a factor $r$. This method is intended to be used with a random path only (unilateral would not make sense). The reduced parameters are:

- The data events size, where $n$ becomes $n'=n/r$.
- The threshold, where $t$ becomes $t'=tr$
- The fraction of the scanned TI, where $f$ becomes $f'=f/r$.

Figure 8 shows simulations using different values for $P_r$. The parameters reduction factor is $r = 2$, which means that when parameters reduction starts, neighborhoods are made of 2 times less nodes, the threshold is doubled and the fraction of TI to scan is half.

The later parameters reduction is started, the better small structures are reproduced. In this case, starting parameters reduction at $P_r = 50\%$ of the simulation does not degrade much the results and provides significant speed-up.

| Fixed parameters: | Varying parameters: |
|---|---|
| • $n = 30$<br>• $t = 0.025$<br>• $\delta = 0.5$<br>• $f = 1$<br>• $r = 2$ | • *$P_r$ ranging from 5% to 100%* |

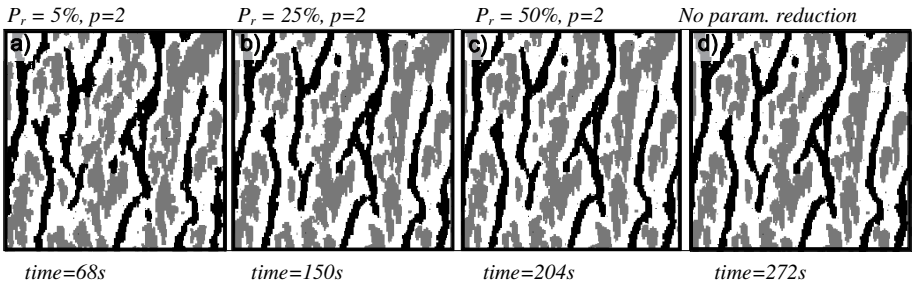



**Figure 8** Effect of parameters reduction on performance and quality

## 5 Detailed description of the Direct Sampling algorithm with the proposed improvements

Note: boxes have been drawn to highlight new parameters and algorithm steps added according to the propositions described above.

**New parameters:**

| | |
|---|---|
| $i_s$ | Max. nb. of non-matching nodes in the data event. |
| $p$ | Number of post-processing passes. |
| $p_f$ | Post-processing factor. |
| $P_r$ | Starting point of parameters reduction. |
| $r$ | Parameters reduction factor. |

**Algorithm steps:**

1. Attribute conditioning data to SG.
2. Loop $p$ times. If it is the second loop, divide $n$ and $f$ with a factor $p_f$.
   a) Define a path in the SG (random or unilateral) that does not include the conditioning data.
   b) Loop on each node **x** of the path:
      i) Check if $P_r$ is reached. If yes, adapt $n$, $t$ and $f$ with a factor $r$.
      ii) Find the neighborhood of **x** made of $n$ nodes, defined by a set of lag vectors **L**.
      *iii)* If no neighbor is found for **x**, randomly take a node **y** in the TI and assign its value $z(\mathbf{y})$ zo $z(\mathbf{x})$.
      iv) Define the data event $\mathbf{N}(\mathbf{x},\mathbf{L})$.
      v) Define the window in the TI where $\mathbf{N}(\mathbf{x},\mathbf{L})$ can fit.
      vi) Randomly draw a location **y** in the scan window for an initial scanning point.
      vii) Define the number of scanning attempts in the TI $s$.
      viii) Loop $s$ times:
         A. Define **y** as the next node in the scan window. If the end of the scan window is reached, restart at the first node of the scan window.
         B. Find the data event $\mathbf{N}(\mathbf{y},\mathbf{L})$.
         C. Compute the distance $d\{\ \mathbf{d}_n(\mathbf{x},\mathbf{L}), \mathbf{d}_n\ (\mathbf{y},\mathbf{L})\ \}$ by looping on each of the $n$ nodes in $\mathbf{N}(\mathbf{y},\mathbf{L})$ and by using an appropriate measure of distance.
         D. As soon as the number of non-matching nodes exceeds $i_s$, interrupt the scan for this node and go to A.
         E. If $d\{\ \mathbf{d}_n(\mathbf{x},\mathbf{L}), \mathbf{d}_n\ (\mathbf{y},\mathbf{L})\ \}$ is the lowest distance for this simulated node **x**, keep $\mathbf{y}_{best}$ in memory.
         F. If $d\{\ \mathbf{d}_n(\mathbf{x},\mathbf{L}), \mathbf{d}_n\ (\mathbf{y},\mathbf{L})\ \}$ is under a certain acceptance threshold $t$, the node is accepted and its value $z(\mathbf{y})$ is attributed to $z(\mathbf{x})$. Otherwise, go to i.
         G. If the number of iterations of the loop A-E has exceeded $s$, the node $\mathbf{y}_{best}$ with the lowest distance is accepted and its value $z(\mathbf{y}_{best})$ is given to $z(\mathbf{x})$.
      ix) End loop
   c) End loop
3. End loop

## 6 Conclusion

This study has lead to a better understanding of the Direct Sampling algorithm. The sensitivity analysis has helped to define the effects of each parameter and to isolate which ones are relevant for the algorithm performance and the quality of the resulting simulations. This preliminary step was mandatory in order to improve the quality/performance ratio.

Three acceleration strategies were successfully implemented. The first one fastens the scan by interrupting it prematurely in “desperate cases”. The second strategy consists in performing a quick simulation that respects the large structures, and then to make additional passes on it to refine the details. The third one consists in reducing less constraining parameters at a certain moment in the simulation. This strategy devotes most of the CPU effort on the first simulated nodes, which are the most important because they condition all the remaining nodes. All three strategies are not exclusive.

Globally, these strategies lead to important speed-ups. The exact value of the speed-up is difficult to estimate, because one should compare the times taken to perform simulations together with the quality of these simulations. But it is realistic to affirm that the implemented improvements can accelerate the algorithm by a factor 2 to 8.

## 7 References

Zhang, T., Pedersen, S., and Mccormick, D. (2008), *Patched path and recursive servo system in multiple-point geostatistical simulations*, paper presented at Geostats 2008, 1-5 Dec. 2008, Santiago, Chile.

# Appendix D

## A Matlab® code for Direct Sampling

*"Un pas vers le moins est un pas vers le mieux"*

*Nicolas Bouvier*
*Le poisson-scorpion*

## 1. Overview

The Matlab® code presented below is a very simplified and schematic implementation of the Direct Sampling (DS) algorithm, described in detail in the body of this thesis. It corresponds to the very first version of the DS, and therefore still has characteristics inherited from traditional MP simulations. For instance, it uses fixed templates that have a box shape, which is made very easy by Matlab's matrix-oriented language. Only the core idea of the DS is implemented, and most features described in the thesis are not present (such as continuous variable simulation, possibility to accommodate flexible data events, multivariate simulation, syn-processing, parallel computing, and so on). Multiple-grids are also not implemented (as they should because the template is fixed), and therefore large structures cannot be well reproduced. The code is presented for demonstration purposes only. The algorithm is oversimplified and it should not be used as such for real applications. Moreover, the algorithm is patented and its commercial usage is illegal without authorization (this does not concern academic research and applications).

The code is extremely short, with just over 100 lines, including comments, generation of the training image and visualization of the results. Removing comments and non-essential parts leaves a code running with only 55 lines. This illustrates the simplicity of the approach.

The first section of the Matlab code is devoted to setting the simulation parameters. The sizes of the training image, simulation grid and template, the distance threshold and the maximum fraction of the TI to scan are defined. Then comes the generation of the TI. This is done by adding in an empty grid circular objects whose location and radius are random. The DS simulation is then launched, and mainly consists of two imbricated loops: the first one on all nodes of the simulated grid (this is the sequential simulation), and the second one on all nodes of the training image (the scan of the TI). The simulation is computationally feasible because the second loop is most of the time quickly interrupted.

Another implementation, written in C++, was used for the examples and applications displayed in this thesis. The C++ version has all the features described in the thesis. Concerning performance, the calculation times are several orders of magnitudes shorter than the Matlab version.

## 2. The Matlab® DS code

```
%% Initialization
clear; home;
%Performs a multiple-points simulation by Direct Sampling.
%The training image is generated using objects.
%Parameters are:
simul_size = [80 80];       %size of the simulation: y x
ti_size = [100 100];        %size of the ti: y x
template = [9 9];           %size of the template: y x
fract_of_ti_to_scan = 0.1; %maximum fraction of the ti to scan
thr = 0.0;                  %threshold position (between 0 and 1)

%% Generation of a training image and display
ti=zeros(ti_size(1),ti_size(2)); [xx,yy]=meshgrid(1:ti_size(1),1:ti_size(2));
for i=1:35
    obj_param=[random('unif',1,100) random('unif',1,100),random('unif',4,8)];
    d=sqrt((xx-obj_param(1)).^2 + (yy-obj_param(2)).^2);
    ti(d <= obj_param(3))=1;
end
figure(1); clf; subplot(1,2,1); colormap gray;
imagesc(ti); title('training image'); axis equal tight xy; drawnow;
H=gca; set(H,'position',[0, 0.25, 0.5 0.5]); tic;

%% DS Simulation
%defining shifts related to the template size
yshift = floor(template(1)/2); xshift = floor(template(2)/2);
%reducing the size of the ti to avoid scanning outside of it
ti_size(1) = size(ti,1)-(2*yshift); ti_size(2) = size(ti,2)-(2*xshift);
%creating empty simulation with a wrapping of NaNs of the size "shift"
simul = nan(simul_size(1)+(2*yshift),simul_size(2)+(2*xshift));
%defining path in ti and in simulation
path_ti = randperm(ti_size(1)*ti_size(2));
path_sim = randperm(simul_size(1)*simul_size(2));
sizesim = size(path_sim,2);
progress = 0; tinod = 0;

%looping simulation nodes
for simnod = 1:sizesim

    progress_current=ceil((simnod*100)/sizesim);
    if progress_current>progress
        progress=progress_current; disp([num2str(progress),' % completed'])
    end

    %find node in the simulation grid
    xsim = ceil(path_sim(simnod)/simul_size(1));
    ysim = path_sim(simnod)-((xsim-1)*simul_size(1));

    %define the point and shifting it to avoid scanning NaNs
    point_sim = [ysim+yshift xsim+xshift];

    %define data event at simulated point
    data_event_sim = simul(point_sim(1)-yshift:point_sim(1)+yshift ,...
        point_sim(2)-xshift:point_sim(2)+xshift);

    %scan the ti
    mindist = inf;  %initial best distance
    tries = 0;      %counter of attempts
    max_scan = size(path_ti,2)*fract_of_ti_to_scan; %max number of scans
```

```
    %reducing the data event to its informed nodes
    no_data_indicator = isfinite(data_event_sim);
    data_event_sim = data_event_sim(no_data_indicator);

    while 1==1 %scan the ti

        tinod = tinod+1; tries = tries+1;
        %if arriving at the end of the path, restart from the beginning
        if (tinod > size(path_ti,2))
            tinod = 1;
        end

        %find the point in the ti
        xti = ceil(path_ti(tinod)/ti_size(1));
        yti = path_ti(tinod)-((xti-1)*ti_size(1));

        %find scanned point and data event
        point_ti = [yti+yshift xti+xshift];
        data_event_ti = ti(point_ti(1)-yshift:point_ti(1)+yshift ,...
            point_ti(2)-xshift:point_ti(2)+xshift);

        %if template is totally unknown, take the first value
        if sum(no_data_indicator(:)) == 0;
            simul(point_sim(1),point_sim(2)) = ti(point_ti(1),point_ti(2));
            break
        end

        %find the data event at this point in the ti
        data_event_ti = data_event_ti(no_data_indicator);

        %evaluate the distance between both data events
        %using another distance here allows using continuous variable
        distance=mean(data_event_sim~=data_event_ti);

        %if distance under threshold, the point is accepted
        if distance <= thr
            simul(point_sim(1),point_sim(2)) = ti(point_ti(1),point_ti(2));
            break
        else
            %check if the distance is under the minimum found so far
            if distance < mindist
                mindist = distance;
                bestpoint = point_ti;
            end
            %if max_scan nodes have been scanned, take the best point found
            if tries > max_scan
                simul(point_sim(1),point_sim(2))=ti(bestpoint(1),bestpoint(2));
                break
            end
        end
    end
end

% remove the "wrapping" of NaNs
toc; simul = simul(yshift+1:end-yshift,xshift+1:end-xshift);

%show simul
subplot(1,2,2); imagesc(simul); title('simulation'); axis equal tight xy;
H=gca; set(H,'position',[0.5, 0.25, 0.5*simul_size(1)/ti_size(1) ...
    0.5*simul_size(2)/ti_size(2)])
```

## 3. Using the code

The code can simply be pasted in the Matlab command window and it will execute. Alternatively, it can be copied in the editor window, where modifications on the parameters can take place. For example, increasing the scanned fraction of the TI dramatically increases simulation time. Increasing the threshold parameter reduces CPU cost, but also decreases simulations quality.

Changing the density and shape of the objects can be done with slight alterations of the code. Using a continuous (for example multiGaussian) TI would necessitate changing the measure of distance when scanning the TI, and adopting one of the distances proposed in the thesis.

## 4. Outputs

Running the code as presented above takes about a minute and produces an image similar to Figure 1, with the generated TI on the left and one unconditional multiple-points simulation on the right.

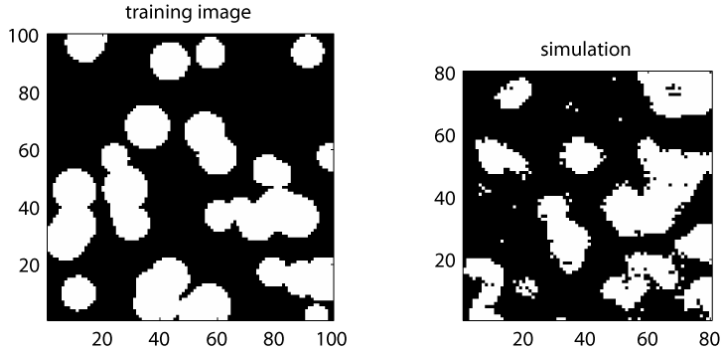



**Figure 1** Training image and simulation produced by the Matlab DS code.

July 2009

Grégoire Mariethoz
Centre d'Hydrogéologie, Université de Neuchâtel
Rue Emile-Argand 11, CH-2009 Neuchâtel
e-mail : gregoire.mariethoz@unine.ch