Computationally efficient geostatistical simulation for uncertainty propagation in models with spatially distributed parameters

Stylianos Liodakis[1],

[1]Department of Geography, University of the Aegean

September 2017

Contents

Abstract

Uncertainty is endemic in geospatial data due to the imperfect means of recording, processing, and representing spatial information. Propagating geospatial model inputs inherent uncertainty to uncertainty in model predictions is a critical requirement in each model's impact assessment and risk-consious policy decision-making. It is still extremely difficult, however, to perform in practice uncertainty analysis of model outputs, particularly in complex spatially distributed environmental models, partially due to computational constraints. In the field of groundwater hydrology, the "stochastic revolution" has produced an enormous number of theoretical publications and greatly influenced our perspective on uncertainty and heterogeneity; it has had relatively little impact, however, on practical modeling. Monte Carlo simulation using simple random (SR) sampling from a multivariate distribution is one of the most widely used family of methods for uncertainty propagation in hydrogeological flow and transport model predictions, the other being analytical propagation. Real-life hydrogeological problems however, consist of complex and non-linear three dimensional groundwater models with millions of nodes and irregular boundary conditions. The number of Monte-Carlo runs required in these cases, depends on the number of uncertain parameters and on the relative accuracy required for the distribution of model predictions. In the context of sensitivity studies, inverse modelling or Monte-Carlo analyses, the ensuing computational burden is usually overwhelming and computationally impractical. These tough computational constrains have to be relaxed and removed before meaningful stochastic groundwater modeling applications are possible.

A computationally efficient alternative to classical Monte Carlo simulation based on SR sampling is Latin hypercube (LH) sampling, a form of stratified random sampling. The latter yields a more representative distribution of model outputs (in terms of smaller sampling variability of their statistics) for the same number of input simulated realizations. The ability to generate unbiased LH realizations becomes critical in a spatial context, where random variables are geo-referenced and exhibit spatial correlation, to ensure unbiased outputs of complex models. On this regard, this dissertation offers a detailed analysis of LH sampling and compares it with SR sampling in a hydrogeological context. Additionally, two alternative stratified sampling methods, here named stratified likelihood (SL) sampling and minimum energy (ME) sampling, are examined (proposed in a spatial context) and their efficiency is further compared to SR and LH in a hydrogeological context; also accounting for the uncertainty related to the particular model at hand via a two step sampling method. All three stratified sampling methods (accounting for model sensitivity in the second case study) were found in this work to be more efficient than simple random sampling.

Additionally, this thesis proposes a novel method for the expansion of the application domain of LH sampling to very large regular grids which is the common case in environmental (hydrogeological or not) models. More specifically, a novel combination of Stein's Latin Hypercube sampling with a Monte Carlo simulation method applicable over high discretization domains is proposed, and its performance is further validated in 2D and 3D hydrogeological problems of flow and transport in a mid-heterogeneous porous media, both consisting of about 1 million nodes. Last, an additional novel extension of the proposed LH sampling on large grids is adopted for conditional high discretized problems.

In this case too, the performance of the proposed approach is evaluated in a 3D hydrogeological model of flow and transport. Results indicate that both extensions (conditional and not) of LH sampling on large grids facilitate efficient uncertainty propagation with fewer model runs due to more representative model inputs.

Overall, it could be argued that all the proposed methodological approaches could reduce the time and computer resources required to perform uncertainty analysis in hydrogeological flow and transport problems. Additionally, since it is the first time that stratified sampling is performed over high discretization domains, it could be argued that the proposed extensions of LH sampling on large grids could be considered a milestone for future uncertainty analysis efforts. Moreover, all the proposed stratified methods could contribute to a wider application of uncertainty analysis endeavors in a Monte Carlo framework for any spatially distributed impact assessment study.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my mentor Prof. Phaedon Kyriakidis for the continuous support of my Ph.D. study and all relevant research conducted during the past five years. I would also like to thank him for patiently transferring his motivation and immense knowledge, encouraging my efforts and for helping me to grow as a scientist. His guidance has been priceless in all the time of research and writing of this thesis, and I certainly feel very lucky for having the chance to cooperate with him all of these years.

Besides Prof. Kyriakidis, I would also like to thank Prof. Petros Gaganis with whom I have closely worked throughout the Ph.D. study. His support has been constant and fruitful along the way and I really feel that he embraced all my efforts and guided me throughout my research on the scientific field of hydrogeology.

I am also really thankful to Prof. Jaime Gomez-Hernandez for his well targeted and constructive criticism to the thesis, which enhanced my final efforts and built a much better overall dissertation.

I would also like to thank Prof. Nikos Soulakellis, for showing trust in my work and accepting to supervise my Ph.D. thesis for this last year due to administrative impediments that did not allowed Prof. Kyriakidis to continue as my supervisor. My sincere thanks also goes to Prof. Kostas Kalampokidis and Prof. Nikos Soulakellis, who provided me an opportunity to join their teams as a student in my Postgraduate diploma and early Pd.D. steps, and gave access to the laboratory and research facilities. Without they precious support it would not be possible to reach at the research maturity necessary to conduct a Ph.D. research.

Additionally, I would also like to thank my fellow labmates Prof. Dimitris Kavroudakis, Nikoletta Koukourouvli and Dr. Giorgos Vasios and for the stimulating discussions and fruitful cooperation we had along the lines of the research program that supported my Ph.D.

A special thanks to my family. Words cannot express how grateful I am to my mother, father, and my brother for all of the support they provided and sacrifices they have made along the way. Your love for me was what kept me going. I would also like to thank my friend and roommate Tegos and my friends Sophia and Antonis for all the unforgettable moments along these years in Mytilene.

At the end I would like to express all my gratitude to my beloved partner Anastasia who really helped me to improve to a better person.

Publicity - Dissemination

This thesis subject matter was funded as part of the research project with title (2269) "Advances in Geostatistics for Environmental Characterization and Natural Resources Management" (GEOSTATENV), implemented within the framework of the Action Aristeia I of the Operational Program "Education and Lifelong Learning" (Actions Beneficiary: General Secretariat for Research and Technology), co-financed by the European Social Fund (ESF) and the Greek State. Along the lines of this thesis outline three deliverables, two interim reports and one final report, were submitted to the research project executive authority and approved.

The following publications were also produced as part of this PhD dissertation:

- In peer - reviewed journals
  - Liodakis S.C., Kyriakidis, P. and Gaganis, P. (2015): Accounting for model sensitivity in controlled (log)Gaussian geostatistical simulation. *Spatial Statistics*, Elsevier, 14(C), 224-239. doi:10.1016/j.spasta.2015.05.007.
  - Liodakis S.C., Kyriakidis, P. and Gaganis, P. (2015): Efficient uncertainty analysis in an anisotropic three dimensional hydrogeological model of flow and transport, *Non Linear Studies / Mathematics in Engineering, Science and Aerospace*, Cambridge Scientific Publishers, 6(4), 657 - 673.
  - Liodakis S.C., Kyriakidis, P. and Gaganis, P. (2017): Conditional Latin hypercube simulation of (log)Gaussian random fields, *Mathematical Geosciences*, Springer. *Accepted for publication.*
- In peer - reviewed conference proceedings
  - Liodakis, S.C., Kyriakids, P. and Gaganis, P. (2015): Efficient uncertainty analysis in a three dimensional hydrogeological model of flow and transport, *In proceedings of 14th International Conference on Environmental Science and Technology CEST2015*, September 3-5, 2015, Rhodes, Greece.
  - Liodakis, S.C., Kyriakids, P. and Gaganis, P. (2015): Efficient uncertainty propagation of lognormal hydraulic conductivity in a three dimensional hydrogeological model of flow and transport on very large regular grids, *In Proceedings of The 17th annual conference of the International Association for Mathematical Geosciences*, September 5-13, 2015, Freiberg (Saxony), Germany.
- In peer - reviewed conference abstracts
  - Kyriakids, P. and Liodakis, S.C. (2012): Latin hypercube sampling of random field models on very large regular grids, 9th Conference on Geostatistics for Environmental Applications (geoENV2012), September 9-12, Valencia, Spain.
  - Liodakis, S.C., Kyriakids, P. and Gaganis, P. (2014): Likelihood-representative sampling from (log)normal random field models, 10th Conference on Geostatistics for Environmental Applications (geoENV2014), July 7 - 11, Paris, France.

- Liodakis, S.C., Kyriakidis, P. and Gaganis, P. (2014): Latin hypercube simulation of hydraulic conductivity fields for efficient parameter uncertainty assessment in flow and transport problems, 10th International Congress of the Hellenic Geographical Society, October 22-24, Thessaloniki, Greece.
- Liodakis, S.C., Kyriakids, P. and Gaganis, P. (2015): Geostatistical sampling methods for efficient uncertainty analysis in flow and transport problems, European Geosciences Union General Assembly 2015, Vienna, Austria, 12-17 April 2015.
- Kyriakids, P., Liodakis, S.C., and Gaganis, P. (2016): Conditional Latin hypercube simulation on large grids, 11th Conference on Geostatistics for Environmental Applications (geoENV2014), July 6 - 8, Lisbon, Portugal.
- Liodakis, S.C., Kyriakids, P. and Gaganis, P. (2016): Conditional Latin hypercube simulation on 3D large grids incorporating uncertainty in model parameters, 10th International Geostatistical Congress (GEOSTATS 2016), September 5-9, Valencia, Spain.

- In Conference Abstracts
  - Liodakis, S.C., Kyriakids, P.C., and Gaganis, P. (2015): Efficient uncertainty assessment methods in spatially distributed hydrogeological models of flow and transport, Meeting at the Technical University of Crete, September, 2015.

Additionally, as part of this thesis I was granted the following two distinctions:

- 'Mathematical Geosciences' Journal student award (2015) for the project "Advances in Geostatistics for Environmental Characterization and Natural Resources Management", $2500.
- Distinction in a Wikimedia Commons competition, in section Non-Photographic Media. A concentration realization from the 3D hydrogeological case study on large grids was assessed in the competition using the Quality image guidelines and is granted as a *Quality image*
  https://commons.wikimedia.org/wiki/File:PollutantConcentration.png

Chapter 1

Introduction

Geostatistics and spatial statistics (Journel and Huijbregts, 1978; Ripley, 1981; Cressie, 1993) is a branch of multivariate statistics that deals with the analysis and modeling of spatial data and has seen significant development over the past thirty years. In contrast to what is assumed in classical statistics, geostatistics handle geospatial data as interdependent, and the phenomenon they describe having a regular variation with regards to space, which is partly known as spatial association/correlation. The lack of independence of observations can create significant inference problems when spatial data are examined by methods of classical statistics (Anselin and Griffith, 1988). Geostatistics provide a comprehensive statistical framework for analyzing and interpreting spatial patterns in geospatial data, for integrating measurements from different sources and resolutions, and for assessing uncertainty in models with spatially distributed parameters. All these tasks are extremely important in scientific research across diverse disciplines (National US Research Council, 1997; National US Research Council, 1998a; National US Research Council, 1998b), such as (hydro)geology (Davis, 1986; de Marsily, 1986) atmospheric sciences (Daley, 1991), oceanography (Emery and Thomson, 1997) and ecology (Zuur et al., 2007).

Modeling is generally the process of representing a real-world object or phenomenon as a set of mathematical equations. Models whose parameters are distributed (vary) in space are called spatially explicit models. The objective in spatial modeling is to be able to study and simulate spatial objects or physical processes that occur in the real world and facilitate problem solving and planning. Modeling of spatial phenomena in earth sciences is performed on a local, regional or global scale, for various scientific or engineering purposes; meteorologists build climate models, geologists build deep earth's structure, hydrologists build subsurface flow models. In any case, the relevant uncertainty analysis should always be part of the overall modeling procedure. The availability of tools in modern science, the huge growth in the use of geographic information systems, remote sensing platforms and spatial databases constitute spatial modeling a major advantage in the hands of researchers, thus more commonly adopted. Spatial modeling can be used for example, to analyze the projected path of tornadoes by layering a map with different spatial data, like roads, houses, the path of the tornado and even its intensity at different points. This allows researchers to determine a tornado's real path of destruction. When juxtaposed with other models from tornadoes that have affected the area, this model can be used to show path correlations and geographical factors. Other examples of spatial models and how they can influence policy making are:

- the modeling of overland flow to predict flood probabilities based on using historical precipitation measurements and a digital elevation model, as well as other factors related to soils and/or geology
- the modeling of ground pollution due to factory discharges and the risk level of reaching the aquifer and endangering the lives of people
- the modeling of animal movement as a result of a year's drought and how it can change the affected landscape
- the modeling of human evacuation behavior used for disaster planning due to, for example, seismic activity
- the modeling of human health risk related to a sudden animal disease outbreak

As it can be easily understood the list of spatial models and their relevance to planning decisions is vast; such models are used in almost every scientific discipline that accounts for spatial data information.

The uncertainty in spatial model results is often a major concern, since it has policy, regulatory, and management implications (Shirmohammadi et al., 2006). Spatial uncertainty is defined as the difference between the contents of a spatial database and the corresponding phenomena in the real world (Goodchild, 2008). Since all spatial models are representations of the real world, it is inevitable that differences will exist between them and the real phenomena. The differences may be due to various sources of uncertainty resulting from input data and parameter variability, model structure, model calibration, spatial and/or temporal scale, model boundary conditions, etc. (Beven, 1989; Haan, 1989; Luis and McLaughlin, 1992), generally categorized into uncertainty related to input parameters and model structure specifications. This thesis mainly focuses on uncertainty related to the input parameters of a spatial model and how this uncertainty is efficiently propagated to spatial model outputs.

1.1 Main objectives

Many efforts have been made, (e.g. Tarantola (2005); Bilcke et al. (2011)), in order to examine uncertainty in spatial models but there is room for further analysis dealing with spatial explicit models and the contribution of spatial uncertainty to optimal sampling design. A holistic uncertainty analysis approach indicates that every single aspect of uncertainty must be carefully appraised and reformulated into ranges of values for each parameter separately. All these ranges should then be compiled into a global model of input parameter uncertainty, and in combination with the uncertainty related to the model parameters should be examined in terms of contribution to the model output uncertainty.

Scientific questions addressed

In this dissertation, an effort is made to contribute to the formulation of answers to the following research questions:

- How to effectively propagate uncertainty related to the inputs of models with spatial distributed parameters to the outputs of such models?
- How to expand the above efficient uncertainty propagation in models operating on highly discretized domains?
- How to apply the above methods when only global statistics (attribute histogram and spatial correlation function) are available (unconditional simulation), and when the location of sample data is also known (conditional simulation)?
- How to apply the proposed methods in a hydreological framework of flow and transport pollutants in heterogeneous porous media, appraising their benefits related to other existing methods?

Expected significance

The findings of this dissertation are expected to contribute to:

- Enhanced uncertainty analysis in outputs of complex models with spatially distributed parameters in highly discretized domains.
- Risk-conscious decision making accounting for uncertainty in model parameters and outputs.
- Development of novel methods for uncertainty analysis in models operating on very large (possibly with million of nodes) regular grids.
- Applications in three dimensional complex hydro(geo)logical models of flow and transport of pollutants in heterogeneous porous media.

1.2 Uncertainty

”Uncertainty factors” are endemic to nearly all environmental characterization and must be accounted for in risk-conscious natural resources management. A wide spectrum of scientific disciplines, such as engineering science (Ang and Tang, 1990), and ecology (Zuur et al., 2007), deal with uncertainty in almost any aspect of data analysis and modeling. Moreover, uncertainty is endemic in geospatial data due to the imperfect means of recording, processing, and representing spatial information (Zhang and Goodchild, 2002). Geospatial data often serve as inputs in environmental models, hence the need for propagating their inherent uncertainty to uncertainty in model predictions.

Limited data, most often unverified and from several different primary sources, are usually available to support spatial analysis. In addition, methods used in the analysis of spatial explicit models (e.g. Dunning et al. (1995)) confront the same uncertainty issues, as they usually involve subjective expert opinion knowledge. Yet outputs of geostatistical analyses are used by managers and relevant agencies in making regulatory decisions. These scientific and technical uncertainty factors should first be studied in order to enable risk-conscious decision making.

Uncertainty analysis is often difficult since it stretches our understanding of the measurements to the limit. In most cases, the lower the uncertainty level required in the analysis the greater the understanding required in terms of both data acquisition and model structure (Nicholas and White, 2001). For this reason, detailed uncertainty analysis has a second purpose: it provides a measure of our competence. This is one of the reasons for emphasizing uncertainty analysis in calibration and testing environments, especially where laboratory accreditation is sought.

Variability and uncertainty

The distinction between uncertainty and variability is critical in environmental modeling. According to Cormack (1988), spatial variability in the measurements of environmental variables is a critically important feature of natural systems, and should not be neglected in modeling. As noted above, variability reflects a type of heterogeneity between individual members of a population, it may refer to several different varying quantitative parameters that characterize individuals, and it is expressed using a frequency distribution. Thus, the frequency distribution of a population acknowledges differences between individuals, hence the term variability. This frequency distribution is used for determining whether the population should be disaggregated into smaller groups that are more homogeneous and should be examined on a case by case basis.

Uncertainty, however, often results from measurement error or other sources as described above. In addition, there is an amount of uncertainty endemic in the construction of a frequency distribution of individuals, i.e., in the distribution itself that characterizes variability. For example, assume a model characterizing the potential amount of water absorbed by the ground, and consider one (of many) input variable, e.g., precipitation, measured at ten meteorological stations. Total model uncertainty, could be partitioned into:

1. Variability, due to the fact that precipitation varies across the study region,
2. Uncertainty, due to the interpolation of precipitation from limited samples, and
3. Overall uncertainty, although simplified in this example, enters into the model due to the small amount of sample data and due to variability in the different distributions of precipitation at each sample location.

Both notions of 'uncertainty and variability' of measurements, contribute to the procedure of statistical analysis and it is desirable to separately characterize them. Uncertainty represents partial ignorance or lack of perfect information about poorly-characterized phenomena or models and calls for further investigation of alternative research procedures and measurement techniques in order to reduce it. Variability represents diversity or heterogeneity in a well characterized population providing information on frequency distributions and recognizes significant subpopulations which merit more focused study.

Variability is usually not reducible through further measurement or study. For example, different people have different body weights, no matter how carefully we measure them (Frey and Burmaster, 1997).

1.2.1 Taxonomy of uncertainty

As mentioned before, environmental model predictions are uncertain due to uncertainty stemming from different sources. Sources of uncertainty have been thoroughly studied by several authors, including Morgan and Henrion (1990), Frey and Cullen (1999), and others. The main concepts involved are highlighted hereafter:

Model uncertainty

The structure of mathematical models used in statistical or environmental analysis is a key source of uncertainty, as they are only a simplified representation of a real-world system, and hence models may be incomplete or incorrect. Building a model requires significant approximations, different scientific or technical assumptions and limited spatial or temporal resolution (e.g., grid size), which may all introduce uncertainty into model results. Different expressions of model uncertainties are summarized according to Frey (1992) as follows:

- *Model Structure*: Alternative sets of scientific or technical assumptions may be available for developing a model. The implications of these alternative foundations may be evaluated by constructing alternative models and comparing results from each alternative model. If similar results are obtained by different competing models that lead to similar decisions, then analysts can be confident that the decision is 'right' even in the face of alternative theories. If, on the contrary, different conclusions are reached, the judgment of an analyst or a decision maker may be required to choose the most suitable decision for a given problem.

- *Model Detail*: Simplifications of complex models (non linear to linear), often adopted for the sake of convenience, induce uncertainty in the outcomes of a model. Simplifications are also made due to lack of knowledge for the basic structure of the actual model.

- *Model Resolution*: A model characteristic that is affected by uncertainty is model resolution. Usually the decision maker has to strike a balance between accuracy and computation time, although the main factor for deciding on the proper grid size is the data resolution of most input variables. Standard techniques are often available to help select the appropriate grid size for reaching the desirable model accuracy. This type of model uncertainty is thus dealt with through the appropriate selection of model domain parameter values, or by comparing results based on different grid sizes. NOTE: Here resolution means parameter resolution, i.e., how many values will one place between 0 and 100 for example.

- *Model Boundaries*: The selection of a models boundaries may be a type of simplification. Reality often sees no boundaries. If, for example, a model is examined within a certain region, the dependent variables may interact with variables outside these boundaries that are not included in the model and hence are ignored.

- *Validation*: Model validation is an indispensable procedure for conducting research whose outcomes might not be placed under controversial dispute. Uncertainties regarding model parameters for which few data are available to test model outcomes may rely on experts opinion in order to evaluate their predictions.

- *Extrapolation*: Even validated models that end up with conclusions for a certain portion of the model's region may be completely inappropriate for making predictions in other regions of the study area or over the whole study area.

- *Scenario Reasonableness*: A set of alternative scenarios must be developed in every model building procedure. For pollutant emissions for example, the analyst should decide on the source, the pathways of pollutant transport and deposition, the populations exposed to the pollutant, and so on. Uncertainty will be introduced to the extent that the scenario fails to consider all factors affecting model outputs. Similar to the uncertainty associated with model boundaries, only the analysts creativity and knowledge of the real possible variables related to the problem can minimize uncertainty.

Parameter uncertainty

Uncertainty of measurable empirical quantities (e.g. precipitation, disease incidents, vegetation), provided as spatial inputs in model predictions is unambiguously inherent in any model (e.g. Hansen et al. (1999); Bennett et al. (2000); Hunsaker et al. (2001); Elith et al. (2002)). Some identifiable sources of uncertainty in empirical quantities are highlighted below according to Morgan and Henrion (1990):

- *Random Error and Statistical Variation*: Imperfect measurement techniques create this type of uncertainty in data. Statistical analysis of test data is thus one method for developing a representation of uncertainty in a variable, often visualized in terms of histograms of attribute values.

- *Systematic Error*: The mean value or variance of a measured quantity may not converge to the "true" corresponding value because of biases in measurements and procedures, such as inaccuracies in the conjectures used to translate actual quantities of interest from proxy variables. For example, there is often systematic error involved in using small scale tests to estimate the values of quantities for large scale systems.

- *Variability*: Some quantities are variable over time, space, or some population of individuals (broadly defined) rather than for any individual event or component. Such variability is rather tricky to be modeled, usually because of the boundaries imposed to the model's analysis and bears a great amount of uncertainty that is up to the analyst to detect.

- *Inherent Randomness or Unpredictability*: Some quantities may be irreducibly random even in principle. This concept of uncertainty is also applied to variables that can be measured precisely, but due to cost or difficulties in calibration they may appear random.

- *Dependence and Correlation*: Multiple uncertain variables in a model may be statistically or functionally dependent. Ignoring this inter-dependence may lead to an incorrect prediction of the output's variance. Dependence among model input variables often arises because of model simplifications which fail to explicitly model the source and form of dependence between them.

1.2.2 Methods of uncertainty analysis

In developing estimates of the values of key quantities in environmental problems, a common approach is to assume a "best guess" point-value of a parameter or variable based on some combination of data and technical judgment. However, the basis for many assumptions, and the scope of thought that went into them, are often not explicitly documented in policy studies. Moreover, the degree of confidence that a decision-maker should place in parameter estimates when evaluating regulatory alternatives is often not rigorously considered.

Uncertainties are often handled by the use of simple "sensitivity" analysis, a technique for systematically changing the values of one or few model input variables at a time in order to determine the effects of such changes in model outputs. In practical problems with a large amount of uncertain input variables, the possible combinations in sensitivity analysis (e.g., one variable "high", another "low," and so on) become unmanageable. Furthermore, sensitivity analysis provides no indication on how each outcome should be weighted and which is the most representative state of reality (Frey, 1992). Sensitivity analysis is a rather deterministic procedure, based on subjective decisions, not providing decision makers with the magnitude and main source of the underlying uncertainties. These deterministic estimations do not provide a complete consideration of interactions and correlations among multiple uncertain variables, which can be rather misleading, especially when dealing with outcomes of complex nonlinear models. According to McCarthy et al. (1995) and Bart (1995), sensitivity and uncertainty analysis are two different procedures that spatially explicit models must be subjected to, as part of their development and testing.

Many facets of uncertainty involved in environmental modeling are usually ignored or seldom considered using sensitivity analysis, since it can only rank parameters according to their relative influence on model predictions. According to Frey (1992), the most appropriate way to analyze and finally delimit uncertainty calls for quantitative approaches. Quantitative estimates of uncertainty, using simple methods, e.g. Monte Carlo techniques, provide decision makers with the real magnitude of uncertainties in each variable separately and cumulatively in the entire model. Dealing with the uncertainty of each variable separately may lead to down-weighting or excluding variables whose uncertainty might be critical for model performance. According to Jager and King (2004), sensitivity analysis may be considered as a preparatory stage of uncertainty analysis, since its purpose is to identify key parameters for more careful measurement in the hope of reducing uncertainty. Moreover, sensitivity analysis, often leads to non-quantitative results, hence the need of a more integrated approach for the assessment of uncertainty.

A more comprehensive approach to uncertainty analysis is the use of probability distributions; see, for example, Kyriakidis (2001). Based on these distributions, statistical simulation techniques provide a more robust insight at simultaneous uncertainties of input variables due to their spatial correlation and can examine their combined effect on model outputs. Statistical simulation provides both the possible range of values for model output parameters, and information about the likelihood of obtaining various model results.

The most common means for assigning parameters (mean and variance) to a probability distribution is to base inference on empirical data. However, in many practical cases, the lack of abundant data relevant to the problem may not suffice for a rigorous statistical analysis. Instead, the analyst might resort to other auxiliary information in order to augment the model's input parameters, and this may lead to additional uncertainties which are difficult to quantify. Thus, available data should be examined in terms of their relevance to the nature of the problem at hand. Additionally, it is well known that even the application of certain statistical procedures requires 'expert opinion ', such as goodness of fit tests, where the analyst makes judgments about what types of parametric distributions should be used for the representation of variability on a given data set (Anderson and Hattis, 1999). Even in Bayesian uncertainty assessment, distributions may be based on empirical data and considerations, such as technically informed distribution selection. People with different information or theoretical beliefs may choose different distributions for the same variable (Morgan and Henrion, 1990).

Linear & Nonlinear models

Linear regression was the first model to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters, and because the statistical properties of the resulting estimators are easier to determine. According to Hoef et al. (2001), statistical models such as linear regression posit simple relationships among the variables that may be unrealistic, but have the virtue that the uncertainty of the estimated parameters in the model can typically be quantified. Following this trend, uncertainty of linear model outputs is easily quantified from the uncertainty of input parameters. With a simple quantitative uncertainty analysis of the model's parameters, we can compute the mean and variance of model outputs

On the contrary, when dealing with nonlinear models, more complicated methods should be employed in order to evaluate total uncertainty in model predictions or outputs. The two most commonly used methods of quantifying uncertainty in non-linear models are

1. **Linearization**: Piece-wise or segmented linear regression is a type of regression in which the independent variable is partitioned into intervals and a separate line segment is fitted to each interval. Segmented regression is useful when the independent variables, clustered into different groups, exhibit different relationships with the dependent variable in these regions. The boundaries between the segments are

breakpoints. Oosterbaan et al. (1990) have presented a method for calculating confidence intervals of the breakpoints so that the breakpoint with the smallest interval, i.e. the optimum break point, can be selected. After segmentation, each separate part of the overall nonlinear pattern can be examined in terms of first-order linear models.

2. **Monte Carlo simulation**: Monte Carlo simulation is also used to quantify uncertainty into complex nonlinear models. As opposed to analytical approximation of model output uncertainty, it consists of generating alternative samples (realizations) from the input parameters, evaluating the model response for each of these realizations, and constructing the corresponding distribution of model predictions. Any realistic uncertainty analysis, however, calls for the availability of a representative distribution of model outputs, and can become extremely expensive in terms of both time and computer resources in the case of complex models (Helton and Davis, 2002). In a spatial context, where variables are spatially correlated, Monte Carlo simulation calls for the Cholesky decomposition of the covariance matrix between variables to generate simple random samples (realizations) based on the mean and variance of the random variables assumed for each point, as well as on the correlations between them; in a spatial context, correlations among random variables are expressed in terms of correlograms and variograms (Helton and Davis, 2002).

1.3 Uncertainty analysis of models with spatially distributed parameters

Uncertainty is the result of the imperfectness of our knowledge when trying to characterise any (spatial or not) phenomenon, It is inherently present in every possible (spatial) model due to potential lack (either in availability or accuracy) of data and more possibly lack of holistic understanding of the underlying phenomena and processes taking place. Uncertainty essentially stems from the fact that it is impossible to characterise the true distribution of a studied property between data. Thus, uncertainty is not an attribute of the studied process itself, and unlike error, it cannot be measured in an absolute way. It is rather an attribute of the interpretation of the available empirical data - it is a model-dependent property. Various tools exist for modeling uncertainty of complex Earth systems, and they should be incorporated in every modelling attempt under the prism of the impact of uncertainty on practical decision problems (Caers, 2011). The importance of analyzing, visualizing and communicating the uncertainties is unquestionable. However, there is no generally accepted unified approach for this task. Ignoring uncertainty can at best lead to slightly incorrect predictions and at worst can be completely fatal regarding the policy making decision and trust which might have been put in the work of the system or operator. Although various books and papers have been written on spatial modeling, spread over many scientific disciplines of earth sciences, the issue of uncertainty is often mentioned just as a side note; it is still rarely used for quantitative and predictive purposes. On the other hand, the usefulness of any model depends in part on the accuracy and reliability of its output. Yet, because all model outputs are subject to imprecision, it is of crucial importance to all users of earth systems modeling that awareness of uncertainty and error should be as widespread as possible. A minimal response to uncertainty awareness should be that users are aware of the possible complications to their analysis caused by uncertainty, and at best deliver the model results with a report of the uncertainty and a variety of plausible outcome influences. An exhaustive uncertainty survey is to present the results of a full modeling exercise which takes into account all types of uncertainty related to the different parameters employed in the analysis (Fisher, 1999).

According to Caers (2011), any modeling of uncertainty is only relevant if made dependent on the particular decision question or practical application for which such modeling is called for. The first step of an exhaustive uncertainty analysis should be to quantify the impact of the relevant uncertainty to the decision problem at hand. There is no particular reason to perform uncertainty analysis of neither the model inputs nor the model specifications without first thoroughly examine the impact it will have on the original question: e.g. do we build a damn in this stream? Is contamination going to reach the aquifer? The possible uncertainty factors in an uncertainty analysis of a real world spatial model could be infinite but there is no added value in their inclusion unless they are related to the analysis objectives. Performing uncertainty analysis taking into account the purpose of it, makes the task of building a model of uncertainty efficient computationally (less time consuming and computer resources needed) and effective in terms of the application envisioned. Another important aspect of the preliminary work for an efficient uncertainty analysis is the inclusion of the scale factor. When assessing uncertainty, the first important decision is the scale of the assessments. Models of local uncertainty are specific to a single location; spatial or regional uncertainty modeling requires uncertainty assessment of attribute values at several locations taken together (Goovaerts, 2001).

This thesis spatial uncertainty analysis is preformed through various hydrogeological case studies of flow and transport in a porous medium. It is a difficult area for modeling uncertainty, since the subsurface is complex, the data are sparse or at best indirect, a medium exists that can be porous and/or fractured. Along the lines of this thesis many simplifications are made in order to focus on methodological approaches that highlight efficient uncertainty analysis of spatial models. My future hope is to further employ these methods in real multi-parametric problems.

Quantitative hydrogeology is considered born in 1856 when Henry Darcy published *The Fountains of the City of Dijon* (Darcy, 1856), which contained the first description of the law governing the flux of water through porous media. The same year Paramelle published a book entitled *The Art of Discovering Ground Water* (Paramelle, 1856) which overcame Darcy's book becoming the best seller of that time. It was analyzing ground water observations in thousands of places in France and developed empirical rules to infer the presence of ground water according to geological and geomorphological observations. For a detailed reference on the history of stochastic hydrogeology the reader is referred to Renard (2007).

This thesis employs quantitative hydrogeoloy in a Monte Carlo context to highlight several proposed novel methods of uncertainty analysis. Uncertainty analysis has a wide range of applications and actually delves and should be accounted for, in every possible effort to model a physical process. Along the lines of this dissertation an excessive literature review on uncertainty analysis in models with spatially distributed parameters (hydrogeological or not) was performed to thoroughly explore all possible aspects/methods/applications. This exploration of different scientific approaches of uncertainty analysis contributes in the general perspective of this work, offering the groundwork upon which this PhD thesis is structured.

Reservoir modeling

Uncertainty analysis in geostatistical reservoir modeling is a necessity when it comes to decision making problems like 'Where should we drill next?' for every possible geological reservoir (water, oil, gas etc.). Many different methods/approaches of estimating geological reservoirs exist and during the last years uncertainty analysis has an important role in every modeling attempt. Caers (2001) employed the use of training images in spatial locations considered jointly. He claimed traditional practice of two point geostatistics for reservoir characterization via the variogram study as incapable of capturing the measure of geological continuity. The proposed method is based on the ability of training images to enhance the model construction with important curvi-linear geological information, beyond the modelling capabilities of the variogram. Through this lens, stochastic reservoir simulation consists of anchoring the borrowed geostructures in the form of multiple-point statistics to the actual subsurface hard and soft data. The proposed method was further investigated in Caers (2011), which can be considered as a milestone in the latest geostatistical uncertainty analysis within a reservoir modeling framework.

Another holistic example of uncertainty and sensitivity analysis in geostatistical reservoir modelling is the one of Kamali et al. (2013). According to this work, "Underground modeling utilizes a wide variety of data, probably in different scales and accuracy to construct reservoir models which are able to represent geological heterogeneities"; quantifying uncertainties in this context produces numbers of possible model output enabling the end user to mitigate the risk of any decision. Modeling of "estimation error" in form of uncertainty analysis is always very important since all geostatistical methods used in estimation of reservoir parameters are inaccurate. In the work of Kamali et al. (2013) sequential gaussian simulation and stochastic models adopting this method is thoroughly reviewed. More specifically, uncertainty quantification of the adopted stochastically populated models is performed and further sensitivity study of the model properties are also quantified. The case studied reservoir is located at carbonate sequences of Sarvak Formation, Zagros, Iran and it comprises of three layers holding different portions of the reserve. Reservoir simulation is clearly defined by the sensitivity analysis on the parameters of stochastic models and petrophysical properties adopted.

Ecological modeling

Spatial uncertainty underlies in all ecological models that are used to understand and manage natural systems, propagates into model predictions, thus hindering any crucial policy making decision. Jager and King (2004) offers a comprehensive overview of uncertainty analysis in ecological modeling. Offering a thorough understanding of why ecological modeling should be performed through spatial models (nature of the problems to be answered, interdependence of various systems with spatial reference in one problem, wide availability of spatial data), the authors further present a classification scheme for questions about the relationship between predictions of an ecological model and variation in its spatial inputs:

1. Uncertainty analysis. How does uncertainty in spatial data influence uncertainty in model predictions?

2. Sensitivity analysis. Which spatially distributed input variables are the model most sensitive to?

3. Error analysis. How do measurement and cartographic errors propagate through the model?

4. Error budget analysis. What sources of error in the processes used to obtain spatial input data cause the largest variation in model predictions?

5. Decision analysis and risk assessment. Given the variation known to exist among realistic alternative input maps, what is the optimal decision (or, alternatively, the ecological risk) predicted by the model?

6. Hypothesis testing using neutral models. What influence does variation in spatial structure have on model predictions?

This classification could be adopted in the mostly multilevel sense of ecological problem, thus offering an analysis framework to reveal relationships between ecological model predictions and spatial uncertainty.

In continuation of the above work, Jager et al. (2005) published a work on uncertainty analysis of spatial population models, which illustrates the ecological consequences of spatial uncertainty for complex multi-parametric landscapes. More specifically this work compiles a detailed overview of habitat population ecological analysis within a Monte Carlo conditional simulation framework. Results show that processes forming variables of attraction for a habitat population mostly control the effects of spatial uncertainty. Complex dynamics generated by combinations of attributes and environments are most sensitive to spatial uncertainty, hence habitat populations residing in chaotic dynamics are more sensitive to slight shifts in demographic parameters forming a particular landscape. The proposed conditional stochastic simulation approach produces lower levels of uncertainty than alternative methods used in landscape ecology since local values of descriptive parameters are taken into account in the spatial modeling. Along these lines the authors refer to their method of efficiently simulating realizations of landscape by the proverb: "To draw an analogy with painting, a random map is non-representational, modern art, whereas each of our maps is an impressionists rendering of the original landscape".

Another proof of the necessity of uncertainty analysis and how it impacts and should predetermine an environmental model analysis is the exhaustive technical report/overview of IPCC (2000) discussing uncertainties in greenhouse gas inventories. The Annex further develops a methodology of estimating inventory uncertainty and sets the following necessary requirements:

- A method of determining uncertainties in individual terms used in the inventory;
- A method of aggregating the uncertainties of individual terms to the total inventory;
- A method of determining the significance of year to year differences and long term trends in the inventories taking into account the uncertainty information;
- An understanding of the likely uses for this information which include identifying areas requiring further research and observations and quantifying the significance of year to year and longer term changes in national greenhouse gas inventories;
- An understanding that other uncertainties may exist, such as those arising from inaccurate definitions that cannot be addressed by statistical means.

The report covers a wide range of the uncertainty analysis potential parameters to be accounted for and pathways to be followed setting the channel for a comprehensive conceptual uncertainty analysis in spatial models of the greenhouse phenomenon and ecological models in general.

A sector of ecological science where uncertainty analysis has an important role in existing literature is landscape classification. The impact of human activity on the structure of the landscape is of great interest in the past decades, especially in areas with a high population pressure where the preservation of certain landscape features are considered important. Landscape monitoring, highlights areas that are more susceptible to changes in land cover, and can propose actions to prevent valuable landscape features from disappearing. Canters et al. (2002) offers a study example aimed at assessing the effects of input uncertainty on the outcome of a raster-based model for structural landscape classification. The proposed model uses a DEM and a land-cover map as input, and defines multiple structural landscape classification types in a Monte Carlo simulation framework, to further classify the effect of the input parameters. Simulation experiments proved uncertainty in land-cover classification mostly affecting the determination of the degree of homogeneity of the landscape, whereas DEM error proved to have much less impact on the determination of the landscape type. Study results indicate that input uncertainty may have a substantial effect on the outcome of classification models. Hence, more attention should be paid to the issue of uncertainty propagation assessment when using classification techniques which are probably more appropriate for the modeling of spatially continuous phenomena.

Further discussion on landscape ecology and how it can be examined through connectivity, as a measure expressing the accessibility of a location to individuals from a source area (Fahrig and Merriam, 1985), is offered in the work of De Genst et al. (2001). This paper studies the potential connectivity of red squirrels in a fragmented landscape, using a buffer operation that takes into account the difficulty of moving through the landscape. This difficulty is quantified by the dispersal capacity which expresses how far an individual is prepared to move from a source area (species dependent), and by the characteristics of the landscape separating the location from a source area which expresses the difficulty an individual experiences when moving through the landscape (dependent on the land-cover type). Uncertainty related to the inputs of the model is quantified in terms of a multivariate statistical classification of remotely sensed data and studied using spatial Monte Carlo simulation. Uncertainty related to the model specifications is examined by adopting deterministic model parameters regarding the dispersal capacity and the landscape effect, and studied using fuzzy set theory. Comparing the outcome of error sensitized models to the observed dispersal activity of squirrels, demonstrates how modeling of uncertainty can help to explain the dispersal activity of red squirrels and more generally how far taking uncertainty into account can improve any landscape connectivity model.

Soil modeling

A well known dilemma of geostatical uncertainty analysis literature is whether one should use stochastic simulation or kriging interpolation techniques. Goovaerts (2001), examined spatial uncertainty in modeling continuous soil attributes through kriging-based and simulation-based techniques. The primary question of this research is whether stochastic simulation is always performing better than the computationally cheaper kriging approach in uncertainly analysis assessment. According to the results there are two cases were stochastic simulation is the best alternative and the relative computational constraints should be put aside. When propagation of uncertainty is performed through multiple-point transfer functions, such as flow simulators, and when the modeling of uncertainty is performed over much larger supports than the measurement supports for attributes that do not average linearly in space. In all other situations which may represent the majority of current applications of geostatistics in soil science a kriging-based model should be the first choice.

Hydrological modeling

Hydrological models have been widely used in the past to provide catchment management with information on the interaction of water, energy and vegetation processes distributed over space and time (Wagener and Gupta, 2005). Computational models can be used to quantify surface and groundwater contributions to streamflow and salt export at catchment scale, and have particular importance with respect to the effect of changes in land-use. In general, hydrological models incorporate many parameters (some statistical and some with physical significance), most of which require measurements from resource-intensive field exercises which are used to calibrate the model by statistical methods. A typical hydrological model consists of a large number of coupled equations describing the direction of water flow and providing predictions of e.g. monthly and annual streamflow or salt deposition. Potential inputs may include the spatial mosaic of climate, soil type, topography, land use, estimates of surface runoff, sub-surface lateral flow, recharge, and potential evaporation. In Benke et al. (2008), various computational approaches were investigated for analyzing the impact of parameter uncertainty on predictions of stream flow for a water-balance hydrological model used. The parameters and associated equations which had greatest impact on model output were determined by combining differential error analysis and Monte Carlo simulation with stochastic and deterministic sensitivity analysis. Their approach aim in the simplification of the model by identifying parameters redundant or insignificant and further examining their influence on the model outputs. More specific aims of this work were (i) to investigate the transfer of uncertainty from designated key parameters to the model output by means of appropriate metrics, (ii) to examine the influence of the shape (skewness) in the parameter distributions on prediction error, and (iii) to conduct sensitivity analysis using both point estimates and parameter distributions.

A very interesting synergism of uncertainty and sensitivity analysis in models with spatially distributed parameters is the one of Crosetto and Tarantola (2001), suited to GIS-based models of any degree of complexity, accepting any type of spatial input data and emphasizing in data error modeling. The authors present two applications of uncertainty and sensitivity analysis in GIS modeling:

- The first application highlights the optimization of the allocation of resources needed for data acquisition, thus optimizing GIS-based models whose output can reliably support the decision process; possible in a low cost strategy, based on numerical simulations on a small prototype of the GIS-based model.

- The second application regards the stage of model building for a new GIS application, offering a novel procedure which supports a scientifically sound choice of the model; it avoids subjective preferences of the modeler.

Last, a comprehensive hydrologic case study of flood forecasting modeling is described, thus putting in practice the theoretical aspects discussed and proving their importance in practical spatial modeling

Hydrogeological modeling

Uncertainty in conceptual model structure and in environmental data is of essential interest when dealing with uncertainty in water resources management. To make quantification of uncertainty possible it is necessary to identify and characterize the uncertainty in geological and hydrogeological data (Nilsson et al., 2006). Uncertainty analysis within a hydrogeological context is the overall thematic umbrella of this dissertation. Many efforts to date have examined uncertainty related to various factors of a hydrogeological model and how in particular they relate to the uncertainty of the relevant model outputs.

Many of the parameters in subsurface flow and transport models cannot be estimated directly at the scale of interest, but can only be derived through inverse modeling. During this process, the parameters are adjusted in such a way that the behavior of the model approximates, as closely and consistently as possible, the observed response of the system under study (Vrugt et al., 2008). According to Zhou et al. (2014), forward model requires specification of a variety of parameters, such as, hydraulic conductivity, storativity and sources or sinks together with initial and boundary conditions. It is rather impossible though to build an exhaustive model through sparse data due to the inherent complexity of the hydrogeological context. Inverse models are used to identify input parameters at unsampled locations by incorporating observed model responses, e.g., hydraulic conductivities are derived based on hydraulic head and/or solute concentration data. Zhou et al. (2014), thoroughly explores and tracks the evolution of the inverse methods over the last decades, mostly within the realm of hydrogeology, revealing their transformation, motivation and recent trends. The same paper offers four concluding features for inverse modeling:

- The goal of inverse problem is not just parameter identification, but also improved predictions.

- Stochastic inverse methods are becoming the trend for the generation of multiple realizations, which will serve to build a model of uncertainty on both parameters and states.

- There is a need for methods that are capable to generate geological models as diverse as possible while matching observed data to ensure that the uncertainty in the predictions is properly captured.

- Multiple sources of observations are integrated in the inverse modeling, multiple scale problems are handled and multiple new algorithms are introduced into the inverse modeling, for instance, multiple point geostatistics and wavelet transform.

Another approach representing attractive alternatives to deal with model error in risk estimation and decision analysis is offered by Gaganis and Smith (2008), comparing two alternative approaches to quantify the conceptual model error in a single groundwater model of 90Sr migration to water wells at Chernobyl, Ukraine. The two methods are a per-datum calibration methodology (Gaganis and Smith, 2006) and a Bayesian model error analysis (Gaganis and Smith, 2001). The comparison of the two methods demonstrate their utility to accounting for the uncertainty due to model error in estimating risks in decision models. Bayesian method for model error quantification resulted in a more conservative estimate of the probability of the Pripyat Town well field contamination than did the per-datum calibration approach. This difference in risk estimates is a result of the conceptual differences between the two methods that result from different weights placed on the probability distributions assigned to uncertain parameters based on prior information. The uncertainty related to model and measurement errors estimated by per-datum calibration is primarily based on information drawn from measurements of the dependent variable. In the Bayesian method, prior information on the parameters and measurements of the dependent variable are effectively equally weighted, thus allowing the incorporation of a more informative description of the parameter space, as well as subjective judgement into a risk analysis. In general, per-datum calibration may be best suited in the case of a large database regarding measurements of the dependent variables, while the Bayesian method may be more suitable when measurements of the dependent variables are limited.

The work of Jang and Liu (2004) evaluated in the Choushui River alluvial fan in Taiwan, using ordinary kriging and sequential Gaussian simulations. A flow model was build and simulated heads using different conductivity realizations were then compared with historically measured heads. Simulated and measured heads were utilized to assess the different spatial variabilities of various estimated conductivity distributions. Evaluating a flow and transport model via Modflow software (also used in this dissertation) showed that ordinary kriging generates realizations with better overall reproduction capability (with lower values of absolute simulation errors) and can be applied to perform a deterministic inverse calibration. Sequential simulation method on the other hand proved a better solution when the application focuses on extreme values of estimated parameters (e.g. in a simulation of contaminant transport in heterogeneous aquifers) and is considered useful in stochastically assessing the spatial uncertainty of highly heterogeneous aquifers.

Simulated annealing is a probabilistic technique for approximating the global optimum of a given function. Specifically, it is a metaheuristic method based on the analogy with the physical annealing process, approximating global optimization in a large search space. Lin et al. (2017) adopted simulated annealing simulation for obtaining limited sets of realizations of hydraulic conductivity of multiple aquifers and spatial correlations among aquifers to simulate realizations of hydraulic heads and quantify their uncertainty in the Pingtung Plain, Taiwan. More specifically, simulated annealing was used to generate large sets of natural logarithm hydraulic conductivity realizations in multiple aquifers based on spatial correlations among them. Moreover, small sets of conductivity realizations were selected from large sets of realizations by ranking the differences among cross-variograms derived from the measured and simulated conductivity realizations between the pairs of aquifers. The uncertainty analysis of simulated realizations of hydraulic conductivity was performed via the generalized likelihood uncertainty estimation. Results indicate that the proposed approach could minimize simulation iterations and uncertainty, and could be effectively applied to evaluate uncertainty in hydrogeological properties and groundwater modeling.

Along the lines of the available literature for stochastic simulation, a great effort was made to try and minimize the computation effort of the demanding iterative simulation process and the relative model output estimation for each simulation at hand. One of the efforts along this way is the one of Jakab (2016), that proposed a distance-based classification of stochastic simulation outputs based on their derived connectivity features. Uncertainty characterization is performed by outlining subsets within the space of output realizations; defining groups that represent characteristically the different model scenarios. To perform the grouping of outcomes, instead of the traditional ranking measures, a static connectivity measure was used as the basis for dissimilarity between the stochastic images. Based on these metrics of dissimilarity, model results holding the same characteristics are easily distinguishable, thus building relationships between output realizations. In that way, one is capable of creating groups of inputs leading to relevant groups of outputs, enabling the selection of representative realizations within these groups that yield a more realistic representation of the smaller scale heterogeneities than all individual stochastic realizations. These advantages of distance based grouping may be of high significance when deriving further reservoir geological or flow properties for petroleum or hydrogeological applications, or spatial patterns of soil or groundwater contaminants. The proposed method offers means to separate the stochastic images representing potential different levels of uncertainties. Additionally, this distance based grouping of outputs is able to calculate the number of realizations that can build an efficient assessment of uncertainty, diminish subjectivity and support reproducible decision-making when the task is to select stochastic realizations to run the full - expensive model.

Other similar attempts aiming at identifying subsets of representative realizations instead of using every single output of the simulation model are those of Scheidt and Caers (2009) in petroleum applications, where the realizations are assigned to subsets based on the dissimilarity distances calculated between them, and Armstrong et al. (2013)

in mining, which is also based on stochastic optimization.

Uncertainty in hydrogeological model predictions is usually related to uncertainty in the hydraulic parameters, whereas according to He et al. (2013), uncertainty in the geological structure should be considered to the same extent. The multiple-point geostatistical method integrated in the Stanford Geostatistical Modeling Software (SGeMS) is used in this work to explore the impact of geological uncertainty on groundwater flow patterns for a particular site in Denmark. Modflow was also used in this work in order to model underwater flow. The relative simulation of groundwater head distribution and travel time were accounted for comparison of three different scenarios. The analysis results indicate that the uncertainty of the conceptual geological model is as significant as the uncertainty related to the embedded hydraulic parameters.

Stratified stochastic simulation methods have also played an important role in uncertainty analysis assessment efforts. The most common stratified simulation method is Latin hypercube sampling, which is thoroughly investigated along this thesis. An example of Latin hypercube application in a hydrogeological context is the work of Dowing et al. (1985). More specifically, in this work Latin hypercube is compared to a sampling method accounting for the functional links between model inputs and outputs; mentioned above as inverse modeling. Latin hypercube method referred by the authors as the "empirical approach" performs iterative sampling from the set of all inputs, evaluates the response surface model for all inputs and obtains the empirical cumulative distribution function of the outputs. The second technique referred by the authors as the "response-surface technique" consists of the following steps: (a) screening to determine the subset of important inputs, (b) response-surface modeling to achieve a proxy to the original code, (c) obtaining moments of the response surface model, and (d) fitting a Pearson or Johnson distribution to the moments to obtain a statistical model of the proxy to the output distribution. The model of dose assessment developed by Hoffman et al. (1982) is employed to determine the efficiency of the two statistical techniques in evaluating model uncertainties. Comparison of the two methods prove the response surface method less adequate and reliable, whereas Latin hypercube is simpler to implement and more reliable estimator of the cumulative distribution function of the model output.

1.4 Contribution of this thesis

Literature dealing with uncertainty analysis prevails in various scientific disciplines and the discussion on methods, procedures and applications is vast. The reference to all the above efforts focuses on highlighting the foreground of major publications that built the theoretical basis of this work and helped in identifying potential gaps that need further evolution and enrichment. Uncertainty quantification in complex spatial models is often performed through least square regression analysis, Bayesian inference approaches, stochastic simulation methods and/or combination of various methods currently available in the literature. On this front, what is identified to be missing is the specialization in stratified stochastic simulation methods, in order to benefit from all the advantages (e.g. efficient uncertainty analysis with minimum computational costs and demanding time) that could derive from the enrichment of existing methods or the proposal of new ones. Along these lines, Chap. 2 presents a detailed overview of Latin hypercube sampling from random fields and how it applies in 2D and 3D hydrogeological problems of flow and transport in a mid-heterogeneous porous media. More specifically, a detailed introduction to simple random (classical Monte Carlo) and Latin hypercube sampling is presented, along with a detailed analysis on how they are applied to univariate, multivariate and random fields problems. Additionally, two hydrogeological case studies are introduced, highlighting the efficiency of Latin hypercube sampling method.

In addition to Latin hypercube, another stratified sampling method is introduced named in this thesis ”Stratified Sampling”. This method is herewith firstly employed in a spatial (hydrogeological) context, and is further evolved to a more efficient sampling version here named ”Minimum Energy Sampling”. Moreover, the proposed methods are also combined with a two step sampling method accounting for the uncertainty related to the parameters of the specific hydrogeological model at hand. These innovative approaches of stratified two step sampling methods could constitute a new proposition for uncertainty analysis via spatial simulation; Chap. 3) offers a detailed overview of all the above.

The overarching objective of this dissertation is to expand the domain of application of Latin hypercube sampling and make it feasible for very large grids. On this regard, Chap. 4 proposes a novel combination of Stein's Latin Hypercube sampling with a Monte Carlo simulation method applicable over high discretization domains and moreover validate its performance in 2D and 3D hydrogeological problems of flow and transport in a mid-heterogeneous porous media, both consisting of about 1 million nodes. The proposed LH sampling method reduces the time and computer resources required to perform uncertainty analysis in hydrogeological flow and transport problems discretized by very large regular grids. Since it is the first time that stratified sampling is performed over high discretization domains, it could be argued that the proposed extension of LH could be considered a milestone for future uncertainty analysis efforts.

Spatial problems recently tend to rely more on spatial data available in the relevant region of each case study, since low cost high detailed data exist due to the constantly improving recording tools. The wide availability and increasing accuracy of data recording tools along with the constantly improving computer resources, play an important role in the number and efficiency of data that can easily (with a relevant low cost) be incorporated in a model. The need to incorporate the available data along the modeling procedure is also considered of great importance, since it is rather more probable to better characterize the entire model by reproducing accurate values in known locations. Along these lines, we further adopt conditional LH sampling on large grids and also evaluate the performance of the proposed approach in a 3D hydrogeological model of flow and transport (Chap. 5).

Summarizing, this work offers novel extensions of stratified sampling in random fields, with demonstrated applications in a hydrogeological context. All the proposed methodological approaches could contribute to a wider application of uncertainty analysis endeavors in any spatially distributed impact assessment study. The proposed methods can, and should be accounted for, potentially in combination with other principles (e.g. inverse modeling), in any future attempt to efficiently quantify uncertainty in any complex environmental model; for a more detailed reference to potential future applications of the proposed methods of this thesis, the reader is referred to Chap. 6.

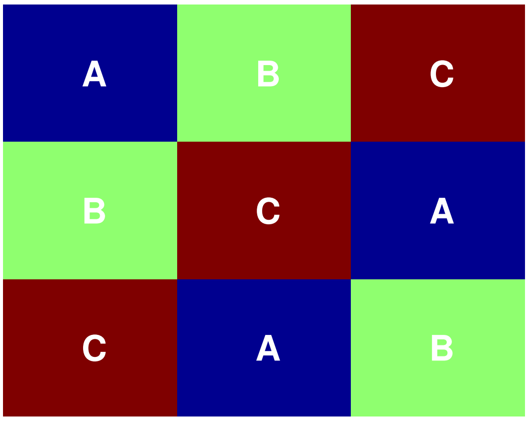
Chapter 2

Latin Hypercube Sampling from Random Fields

Monte Carlo simulation, or simple random (SR) sampling from a univariate or multivariate probability distribution, is routinely used for uncertainty and sensitivity analysis of model predictions in a wide spectrum of scientific disciplines, such as engineering science, hydrology, and more generally earth sciences, to name but a few. Monte Carlo simulation consists of generating alternative samples (realizations) from the input parameters, evaluating the model response for each of these realizations, and constructing the corresponding distribution of model predictions. In a spatial context, the spatial distribution of geo-referenced variables is typically modeled within a geostatistical framework via a random field; that is, a set of spatially correlated random variables, one per location (Chilès and Delfiner, 1999). In hydrogeological investigations involving flow and transport in heterogeneous porous media, for example, the spatial distribution of saturated hydraulic conductivity is often parameterized in terms of a lognormal random field model. Realizations of such a random field are then used along with physically-based simulators of flow and transport in a Monte Carlo framework for evaluating, for example, the uncertainty in the spatial distribution of solute concentration due to the uncertainty in the spatial distribution of hydraulic conductivity and possibly other relevant variables (Gutjahr and Bras, 1993).

Any realistic uncertainty analysis, however, calls for the availability of a representative distribution of model outputs, and can quickly become expensive in terms of both time and computer resources in the case of complex models (Helton and Davis, 2002). This problem is far more pronounced in earth and environmental science applications, where, in hydrogeology for example, three dimensional grids of hydraulic conductivity values are used along with other parameters to simulate flow and transport in porous media (Gutjahr and Bras, 1993). An efficient alternative to classical Monte Carlo simulation is Latin hypercube (LH) sampling, a form of stratified random sampling, aiming at generating representative samples or realizations from a set of random variables with a given multivariate probability distribution. Here, the term representative implies realizations spanning efficiently the range of possible attribute realizations corresponding to that probability distribution.

Latin Hypercube is a statistical method for generating a near-random sample of parameter values from a multidimensional distribution. It is adopted in many scientific papers (e.g. Urban and Fricker (2010), Minasny and McBratney (2006), Olsson and Sandberg (2002), to name but a few) dealing with various disciplines and spatial or not models. The sampling method first described by McKay et al. (1979), is often used for Monte-Carlo integration. The name originates from Latin square; in experimental design Latin square is an $n \times n$ array filled with n different symbols, each occurring exactly once in each row and exactly once in each column. An example of a 3×3 Latin square is depicted in Fig. 2.1. The name ”Latin square” was inspired by mathematical papers by Euler (1782) who used Latin characters as symbols. A Latin hypercube is the generalisation of this concept to an arbitrary number of dimensions, whereby each sample is the only one in each axis-aligned hyperplane containing it.



A	B	C
B	C	A
C	A	B

Figure 2.1: Example of 3×3 latin square further colored for better visualization.

LH sampling has been shown to lead to model outputs with smaller sampling variability in their statistics than SR sampling for the same number of input simulated realizations; that efficiency, however, decreases the more non-linear that model becomes in the parameters (McKay et al., 1979; Helton and Davis, 2003; Helton et al., 2006).

The most widely used methods for generating LH samples from a multivariate distribution are those of Iman and Conover (1982) and Stein (1987). In the first method, the entries of an uncorrelated SR or LH sample are re-arranged to match a target rank correlation matrix. In the second method, a correlated SR sample is transformed into a correlated LH sample based on the ranks of the former; correlation is inherited in the LH sample from the correlation in the ranks of the original SR sample. These methods do not rely on any Gaussian assumption, and both can be used for simulation with or without conditioning data. Relevant representative applications in a spatial context include the work of Zhang and Pinder (2003) and Pebesma and Heuvelink (1999), respectively.

This chapter describes in detail Latin Hypercube sampling in a spatial context further comparing its efficiency to Simple Random sampling in a hydrogeological case study. More specifically, the remainder of the chapter is structured as follows: Section 2.1 describes Latin hypercube sampling from a univariate distribution, Section 2.2 illustrates the extension of Latin hypercube sampling to the general multivariate case, Section 2.3 introduces the concept of a random field and describes algorithms for Latin hypercube sampling of Gaussian random fields, and Section 2.4 presents two synthetic case studies involving flow and transport in a heterogeneous porous medium illustrating the benefits of Latin hypercube over simple random sampling.

2.1 Latin hypercube sampling: Univariate case

In this Section, the concepts of simple random and Latin hypercube sampling are introduced as means for generating realizations (samples) from a single random variable (univariate case). More precisely, Subsection 2.1.1 introduces simple random sampling, Subsection 2.1.2 presents Latin hypercube sampling, whereas Subsection 2.1.3 provides a performance comparison for the two sampling methods.

2.1.1 Simple random sampling

Consider the general case of a random variable (RV) Y with arbitrary cumulative distribution function (CDF) $F_Y(y)$. A simple random (SR) sample of size S can be readily generated from the RV Y via the inversion method, as (Kroese et al., 2011)

$$\mathbf{y} = F_Y^{-1}(\mathbf{u}) \quad \text{or} \quad y_s = F_Y^{-1}(u_s),\ s = 1, \ldots, S \tag{2.1}$$

where $\mathbf{y} = [y_s, s = 1, \ldots, S]^T$ is a $(S \times 1)$ vector with S realizations (simulated quantiles) from Y, $F_Y^{-1}(\cdot)$ denotes the inverse CDF or quantile function of Y, and $\mathbf{u} = [u_s, s = 1, \ldots, S]^T$ is a $(S \times 1)$ vector of uniformly distributed random numbers within the $[0, 1]$ interval. Essentially, y_s is the quantile of Y associated with the probability u_s.

For a RV Y with a parametric Gaussian distribution with mean μ_Y and standard deviation σ_Y, i.e., $y \sim \mathcal{G}(\mu_Y, \sigma_Y^2)$, Equation (2.1) becomes

$$\mathbf{y} = G^{-1}(\mathbf{u}, \mu_Y, \sigma_Y) \quad \text{or} \quad y_s = G^{-1}(u_s, \mu_Y, \sigma_Y),\ s = 1, \ldots, S \tag{2.2}$$

where $G^{-1}(\cdot)$ denotes the inverse Gaussian CDF (Fig. 2.2).

Of particular interest for many environmental applications is the case of a lognormal RV $Z \sim \mathcal{L}(\mu_Z, \sigma_Z^2)$ with mean μ_Z and variance σ_Z^2, linked to a Gaussian RV $Y = \log(Z)$ $\sim \mathcal{G}(\mu_Y, \sigma_Y^2)$ with mean μ_Y and variance σ_Y^2. Due to the functional link $Z = \exp(Y)$, the statics of RV Z are linked to those of RV Y (Chilès and Delfiner, 1999)

$$\mu_Z = \exp(\mu_Y + \sigma_Y^2/2) \quad \text{and} \quad \sigma_Z^2 = \exp(2\mu_Y + \sigma_Y^2)(\exp(\sigma_Y^2) - 1) \tag{2.3}$$

and a SR sample of size S from a lognormal RV Z can then be generated as

$$\mathbf{z} = \exp(\mathbf{y}) = \exp\left(G^{-1}(\mathbf{u}; \mu_Y, \sigma_Y)\right) \tag{2.4}$$

where $\mathbf{z} = [z_s, s = 1, \ldots, S]^T$ is vector with S lognormal deviates (Fig. 2.2).

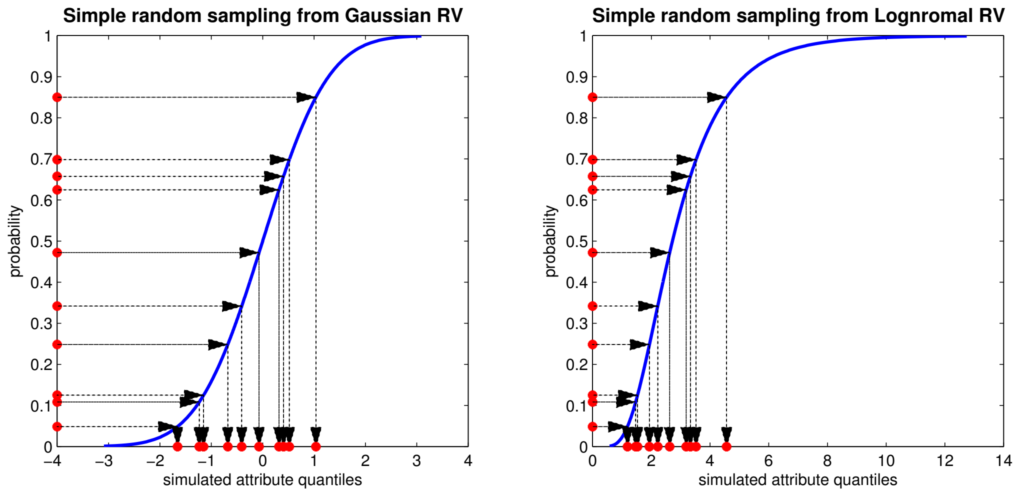


Figure 2.2: Example of simple random sampling of size $S = 10$ from a standard Gaussian RV (left) and a lognormally distributed RV with mean 1 and standard deviation 0.5 (right).

2.1.2 Latin hypercube sampling

Latin hypercube (LH) sampling of S realizations or samples from Y is accomplished by stratifying the probability axis in S equal intervals or strata, generating S stratified random numbers, one within each probability stratum, and computing the corresponding S quantiles from the inverse CDF of Y (McKay et al., 1979)

$$\mathbf{y}_L = F_Y^{-1}(\mathbf{u}_L) = F_Y^{-1}\left(\frac{\mathbf{i} - \mathbf{u}}{S}\right) \tag{2.5}$$

where $\mathbf{y}_L = [y_s^L, s = 1, \ldots, S]^T$ is a $(S \times 1)$ vector with LH realizations (simulated stratified quantiles) from Y, $\mathbf{u}_L = [u_s^L, s = 1, \ldots, S]^T$ is a $(S \times 1)$ vector of stratified random numbers within the interval $[0, 1]$, and $\mathbf{i} = [i_s, s = 1, \ldots, S]^T$ is a $(S \times 1)$ vector of random permutations of S integers $\{1, 2, \ldots, S\}$ generated independently of $\mathbf{u}$.

A monotonic transformation of the stratified probability values in $\mathbf{u}_L$, such as that incurred by the inverse CDF $F_Y^{-1}(\cdot)$, does not ruin stratification; this entails that each entry of vector $\mathbf{y}_L$ (each simulated quantile) falls within a different stratum in the original variable space, no matter the distributional form of $F_Y(y)$. The independence of vectors $\mathbf{i}$ and $\mathbf{u}$ ensures that there is a uniform probability $1/S$ for a simulated value y within any two consecutive quantiles corresponding to a fixed probability interval.

A LH sample of size S from a Gaussian RV Y can be generated as

$$\mathbf{y}_L = G^{-1}(\mathbf{u}_L, \mu_Y, \sigma_Y) \tag{2.6}$$

where $\mathbf{y}_L$ now contains S stratified Gaussian deviates generated via the inverse Gaussian CDF $G^{-1}(\cdot)$; see Fig. 2.3.

Similarly, a LH sample can be generated from a lognormal RV Z as

$$\mathbf{z}_L = \exp(\mathbf{y}_L) = \exp\left(G^{-1}(\mathbf{u}_L; \mu_Y, \sigma_Y)\right) \tag{2.7}$$

where $\mathbf{z}_L = [z_s^L, s = 1, \ldots, S]^T$ is of S stratified lognormal deviates (Fig. 2.3).

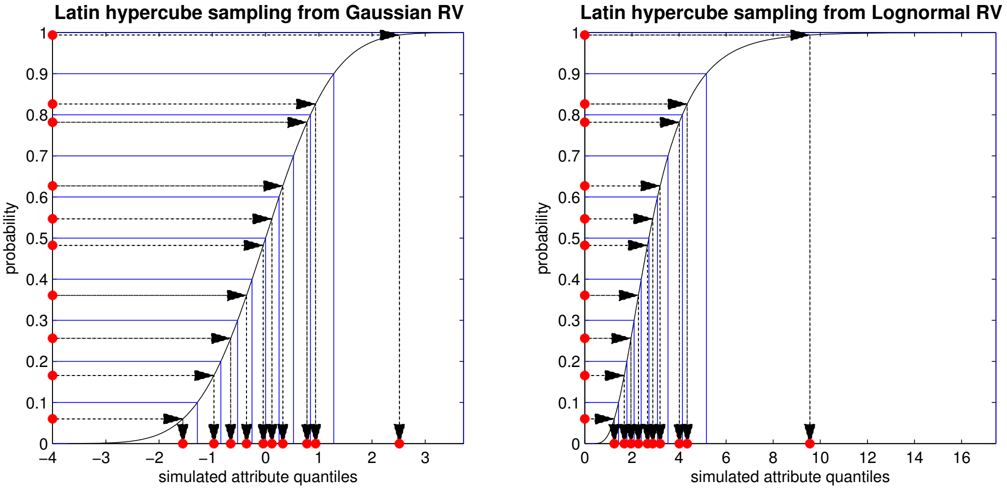


Figure 2.3: Example of Latin hypercube sampling of size $S = 10$ from a standard Gaussian RV (left) and a lognormally distributed RV with mean 1 and standard deviation 0.5 (right).

Although LH sampling yields stratified simulated quantiles, simulated values can still be similar to one another if they lie at the edges of strata. To reduce this effect, one could draw random numbers from a triangular distribution $Tri(0, 0.5, 1)$ in the $[0, 1]$ interval centered at 0.5 within each probability stratum; this preferentially simulates random numbers near the strata centers (Fig. 2.4). Such triangularly distributed random numbers can be generated by adding uniform random numbers in $[0, 0.5]$ (Ang and Tang, 1990; Kroese et al., 2011). At the limit, one could select the S midpoints of the probability strata, hence the S midpoints (quantiles) of the associated intervals in the original variable space; this LH sampling version is termed *midpoint* or *lattice* sampling. In this case, $u_s = 0.5, \forall s$, and vector $\mathbf{u}$ contains S replicates of 0.5 (Fig. 2.4. Midpoint LH sampling, however, might be considered too deterministic, as it amounts to re-using the same quantiles in different order; this might be undesirable for certain applications where extreme values are particularly consequential and simulation involves few samples (small S). Variations of the above basic LH sampling procedure to further control sampling variability include variance reduction techniques, such as antithetic and control variates, as well as correlated sampling (Ang and Tang, 1990; Kroese et al., 2011).

2.1.3 SR and LH comparison

As stated in the introduction, LH sampling has been shown to lead to model outputs (functions of model inputs) with smaller sampling variability in their statistics than SR sampling for the same number of input simulated realizations (McKay et al., 1979; Helton and Davis, 2003; Helton et al., 2006). In what follows, the relative performance of the two sampling methods is evaluated in terms of the reproduction of the known parameters of a standard Gaussian univariate distribution $\mathcal{G}(0, 1)$. That reproduction is evaluated via the sampling distribution of the mean and standard deviation computed from 10,000 sampling repetitions using SR and LH sampling from a standard Gaussian RV, each

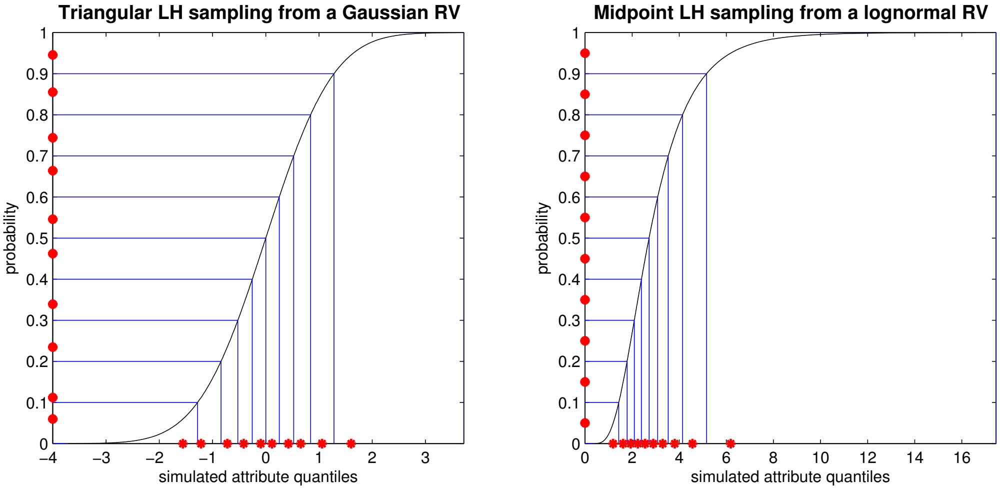


Figure 2.4: Example of midpoint or lattice sampling of size $S = 10$ a standard Gaussian RV (left) and triangular LH sampling from a lognormal RV (right).

sample containing $S = 10, 25, 50$ values. Sampling distributions are presented in terms of their means and medians, as well as their 75% and 95% probability intervals. The better the reproduction of a reference parameter from simulations of a sampling method with a given sample size, the narrower the sampling distribution of the corresponding statistic, and the closer the center of that distribution to the reference parameter.

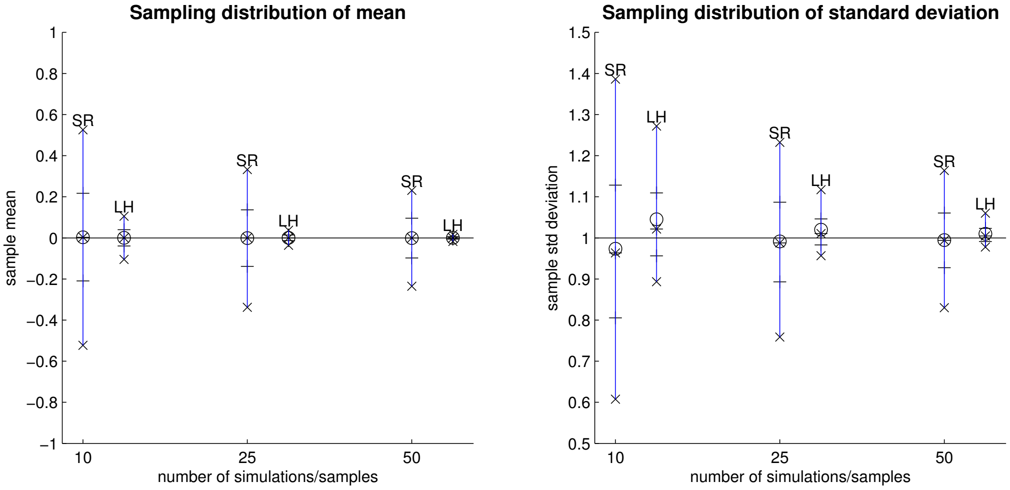


Figure 2.5: Quantiles of the sampling distributions of mean (left) and standard deviation (right) obtained via SR and LH sampling from a standard Gaussian RV with different sample sizes, $S = 10, 25, 50$; see text for details.

More precisely, Fig. 2.5 summarizes the sampling distribution of the mean and standard deviation computed via SR and LH sampling from a standard Gaussian RV using different sample sizes. 75% probability intervals are depicted with horizontal line segments, whereas 95% probability intervals with × symbols; median values are depicted as asterisks ($*$), whereas mean values as circles ($\circ$). Horizontal lines running across each plot depict the reference parameters $\mu_Y = 0$ (right) and $\sigma_Y = 1$ (left). It can be seen that LH sampling leads to much narrower probability intervals than SR sampling, particularly for the case of the sample mean. The efficiency of LH over SR is smaller for the sample standard deviation, since that statistic is non linear in the data values.

A similar illustration is given in Fig. 2.6 for the case of the probability not exceeding two thresholds, namely 0.675 and 1.645 corresponding to the 0.75 and 0.95 quantiles of a standard Gaussian RV. Horizontal lines running across each plot depict the reference parameters $p_1 = 0.75$ (right) and $p_2 = 0.95$ (left). Note again, that LH sampling is more efficient that SR sampling for the same number of samples; that efficiency, however, decreases the more extreme the threshold becomes.

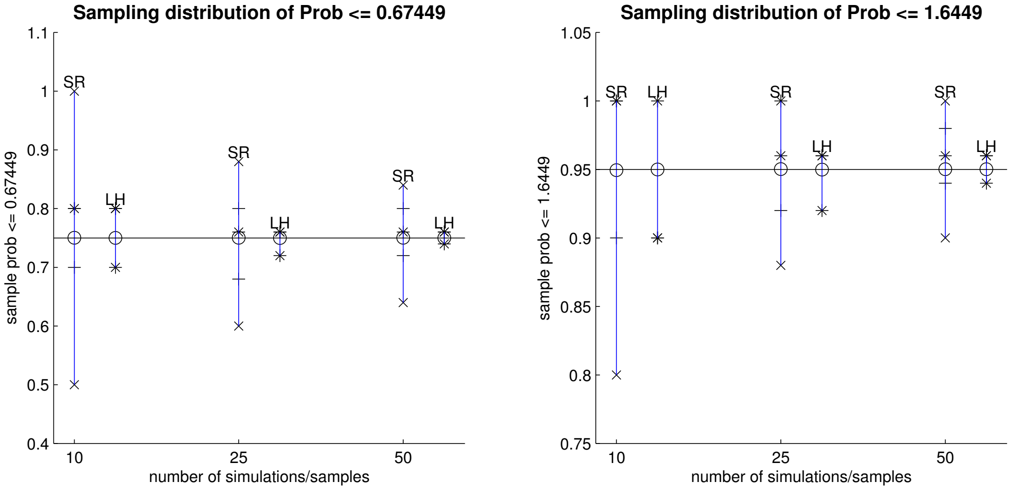


Figure 2.6: Quantiles of the sampling distributions of the probability of not exceeding 0.675 and 1.645 obtained via SR and LH sampling from a standard Gaussian RV with different sample sizes, $S = 10, 25, 50$; see text for details.

2.2 Latin hypercube sampling: Multivariate case

In a multivariate context, one defines M random variables (RVs) comprising a $(M \times 1)$ random vector $[Y_m, m = 1, \ldots, M]^T$, where Y_m denotes the m-th RV whose realization is y_m. Notation-wise, $\mathbf{y}$ will be used to denote both a random vector and its realization; the distinction between the two will be evident from the context of the discussion. The multivariate probability density function (PDF) of the M RVs comprising the random vector is denoted as $f_Y(\mathbf{y})$, and in the Gaussian case it is assumed to be M-variate Gaussian. Consequently, the random vector $\mathbf{y}$ is assumed to follow a multivariate Gaussian PDF $\mathbf{y} \sim \mathcal{G}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$, fully parameterized by the $(M \times 1)$ expectation vector $\boldsymbol{\mu}_Y = [\mu_m^Y, m = 1, \ldots, M]^T$, and the $(M \times M)$ pairwise covariance matrix

$\mathbf{\Sigma}_Y = [\sigma^Y_{m,m'}, m = 1, \ldots, M, m' = 1, \ldots, M]$, where $\sigma_{mm'} = \mathbb{C}ov\{Y_m, Y_{m'}\}$ denotes the covariance between two RVs Y_m and $Y_{m'}$. In what follows, we assume without loss of generality a zero mean expectation vector $\boldsymbol{\mu} = \mathbf{0}$, where $\mathbf{0}$ denotes a $(M \times 1)$ vector with zero entries.

In what follows, Subsection 2.2.1 presents methods for simple random sampling from a multivariate distribution, whereas Subsection 2.2.2 illustrates the application of Latin hypercube sampling in a multivariate context.

2.2.1 Simple random sampling

Let $\mathbf{w} = [W_m, m = 1, \ldots, M]^T$ denote a standard Gaussian random vector with a $(M \times M)$ identity covariance matrix $\mathbf{I}$, i.e., $\mathbf{w} \sim \mathcal{G}(\mathbf{0}, \mathbf{I})$, where W_m denotes a standard (zero-mean, unit-variance) Gaussian RV. The above random vector $\mathbf{w}$ can be transformed into a random vector $\mathbf{y}$ with covariance matrix $\mathbf{\Sigma}$ as (Johnson, 1987):

$$\mathbf{y} = \mathbf{H}\mathbf{w} \tag{2.8}$$

where $\mathbf{H}$ denotes a $(M \times M)$ transformation matrix satisfying $\mathbf{H}\mathbf{H}^T = \mathbf{\Sigma}$.

Indeed, one can easily show that the resulting random vector $\mathbf{y}$ has the desired covariance matrix $\mathbf{\Sigma}$

$$E\left\{\mathbf{y}\mathbf{y}^T\right\} = E\left\{\mathbf{H}\mathbf{w}(\mathbf{w}^T\mathbf{H}^T)\right\} = \mathbf{H}E\left\{\mathbf{w}\mathbf{w}^T\right\}\mathbf{H}^T = \mathbf{H}\mathbf{I}\mathbf{H}^T = \mathbf{\Sigma}$$

where $E\{\cdot\}$ denotes the expectation operator.

There exist multiple definitions of matrix $\mathbf{H}$ satisfying $\mathbf{H}\mathbf{H}^T = \mathbf{\Sigma}$, the most computationally efficient being derived from the Cholesky decomposition of the covariance matrix $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T \tag{2.9}$$

hence $\mathbf{H} = \mathbf{L}$, with $\mathbf{L}$ being a $(M \times M)$ lower triangular matrix termed the Cholesky factor of $\mathbf{\Sigma}$.

An alternative definition of matrix $\mathbf{H}$ is derived from the spectral decomposition of matrix $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \mathbf{Q}\mathbf{E}\mathbf{Q}^T = (\mathbf{Q}\sqrt{\mathbf{E}})(\sqrt{\mathbf{E}}\mathbf{Q}^T) \tag{2.10}$$

hence, $\mathbf{H} = \mathbf{Q}\sqrt{\mathbf{E}}$, where $\mathbf{Q}$ is a $(M \times M)$ orthonormal matrix with $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, hence $\mathbf{Q}^T = \mathbf{Q}^{-1}$, whose m-th column $\mathbf{Q}(\cdot, m)$ is the m-th eigenvector of $\mathbf{\Sigma}$, and $\mathbf{E}$ is a $(M \times M)$ diagonal matrix of eigenvalues of $\mathbf{\Sigma}$, i.e., $\mathbf{E} = diag(\mathbf{e})$, where $\mathbf{e} = [e_m, m = 1, \ldots, M]^T$ is a $(M \times 1)$ vector of eigenvalues; similarly, $\sqrt{\mathbf{E}} = diag(\sqrt{\mathbf{e}})$.

A third definition of the transformation matrix $\mathbf{H}$ can be derived from the square root decomposition of the covariance matrix $\mathbf{\Sigma}$, an extension of the spectral decomposition of Eq. (2.10)

$$\mathbf{\Sigma} = \mathbf{Q}\sqrt{\mathbf{E}}\sqrt{\mathbf{E}}\mathbf{Q}^T = \mathbf{Q}\sqrt{\mathbf{E}}(\mathbf{Q}^T\mathbf{Q})\sqrt{\mathbf{E}}\mathbf{Q}^T = (\mathbf{Q}\sqrt{\mathbf{E}}\mathbf{Q}^T)(\mathbf{Q}\sqrt{\mathbf{E}}\mathbf{Q}^T) \tag{2.11}$$

hence, $\mathbf{H} = \mathbf{Q}\sqrt{\mathbf{E}}\mathbf{Q}^T = \sqrt{\mathbf{\Sigma}}$. Matrix $\sqrt{\mathbf{\Sigma}}$ is a $(M \times M)$ square and symmetric matrix termed the *square root* of the covariance matrix $\mathbf{\Sigma}$, since $\sqrt{\mathbf{\Sigma}}\sqrt{\mathbf{\Sigma}} = \mathbf{\Sigma}$.

Computing the eigenvectors and eigenvalues of an arbitrary covariance matrix $\mathbf{\Sigma}$, however, is a computationally expensive operation, even more so than the Cholesky decomposition of $\mathbf{\Sigma}$. This has led to the Cholesky decomposition being the most widely used method for simulating realizations from a multivariate Gaussian random vector.

Simulation: Simple random (SR) sampling from a, zero-mean, multivariate Gaussian random vector $\mathbf{y} \sim \mathcal{G}(\mathbf{0}, \boldsymbol{\Sigma})$ amounts to generating a $(S \times M)$ matrix $\mathbf{Y}_R = [\mathbf{y}_m^R, m = 1, \ldots, M]$, whose m-th column holds a SR sample of size S from the m-th RV Y_m; the (column-wise) mean of $\mathbf{Y}_R$ is approximately $\mathbf{0}$ and the pairwise (column-to-column) covariance approximates $\boldsymbol{\Sigma}$. Due to computational reasons, the generation of matrix $\mathbf{Y}_R$ is most often performed using the Cholesky factor $\mathbf{L}$ of the covariance matrix $\boldsymbol{\Sigma}$

$$\mathbf{Y}_R = \mathbf{W}_R \mathbf{L}^T \tag{2.12}$$

where $\mathbf{W}_R$ is a $(S \times M)$ matrix of uncorrelated standard Gaussian deviates generated via SR sampling, and $\mathbf{L}^T$ is a $(M \times M)$ upper triangular matrix. Fig. 2.7 gives an application example of Eq. 2.12 for the generation of a sample of size $S = 300$ from a set of $M = 3$, zero-mean, Gaussian RVs with pairwise correlations $\rho_{12} = 0.7$, $\rho_{13} = 0.5$, $\rho_{23} = 0.3$, and st. deviations $\sigma_1 = 4$, $\sigma_2 = 5$ and $\sigma_3 = 6$.

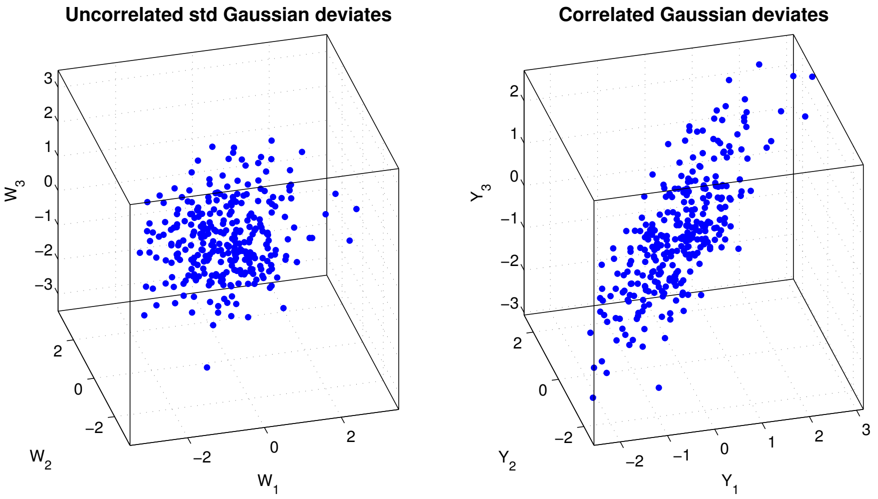


Figure 2.7: A sample of size $S = 300$ from three uncorrelated standard Gaussian RVs (left) and three correlated Gaussian RVs with zero mean and non-identity covariance matrix (right); see text for details.

Multi-lognormal case: Under the multivariate lognormal model, one can generate a SR sample $\mathbf{Z}_R$ from a M-variate lognormal distribution $\mathcal{L}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$ by transforming a SR sample $\mathbf{Y}_R$ generated from a multivariate Gaussian distribution $\mathcal{G}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$

$$\mathbf{Z}_R = \exp(\mathbf{Y}_R) \tag{2.13}$$

where the entries of the lognormal mean vector $\boldsymbol{\mu}_Z$ and covariance matrix $\boldsymbol{\Sigma}_Z$ are derived using Eq. (2.3). In what follows, every sample generated from a Gaussian distribution will be converted in a lognormal one via the above transformation.

2.2.2 Stein's method for Latin hypercube sampling

Latin hypercube (LH) sampling from M random variables amounts to generating a $(S \times M)$ matrix $\mathbf{Y}_L = [\mathbf{y}_m^L, m = 1, \ldots, M]$, whose m-th column holds a LH sample for the m-th RV Y_m. In other words, LH sampling amounts to generating a *marginally* stratified sample (of size S) from M correlated RVs.

A widely used method for multivariate LH sampling is that of Stein (1987), in which a SR sample matrix $\mathbf{Y}_R = [\mathbf{y}_m, m = 1, \ldots, M]$, generated for example via Eq. (2.12), is transformed into a LH sample matrix $\mathbf{Y}_L$. In the Gaussian case, the SR sample $\mathbf{y}_m$ for the m-th RV Y_m, i.e., the m-th column of matrix $\mathbf{Y}_R$, is transformed into a LH sample $\mathbf{y}_m^L$ for that RV, as

$$\mathbf{y}_m^L = G^{-1}\left(\frac{\text{rank}(\mathbf{y}_m) - \mathbf{v}_m}{N}\right) \tag{2.14}$$

where $\mathbf{v}_m$ is a $(S \times 1)$ vector of uniform random numbers in $[0, 1]$, independent of $\mathbf{y}_m$, and $\text{rank}(\mathbf{y}_m) = \mathbf{r}_m$ denotes a $(S \times 1)$ vector of integers from 1 to S corresponding to the ranks of the entries of the SR sample $\mathbf{y}_m$: the smallest simulated y-value for the m-th RV Y_m is assigned a rank of 1, whereas the largest a rank of S.

The LH sample matrix $\mathbf{Y}_L$ contains (column-wise) correlated entries, due to the correlation present in the SR sample matrix $\mathbf{Y}_R$. In addition, the entries of any column of matrix $\mathbf{Y}_L$ are stratified, as opposed to the entries of any column of matrix $\mathbf{Y}_R$. In essence, Stein's method introduces correlation in the univariate stratification procedure of Eq. (2.5), by replacing the array $\mathbf{p}$ of random permutations for the m-th RV Y_m by the array $\mathbf{r}_m = \text{rank}(\mathbf{y}_m)$ of SR sample ranks for that RV (Eq. 2.14).

Fig. 2.8 gives an example of a LH sample of size $S = 10$ from $M = 2$ standard Gaussian RVs with correlation $\rho_{12} = 0.75$, along with the corresponding SR sample; see Eq. (2.14). The vertical and horizontal lines correspond to deciles of a standard Gaussian RV and delineate strata of equal probability. It can be readily seen that the LH sample is marginally stratified; that is, viewed from either the abscissa or ordinate axis, each stratum contains one simulated value. Evidently, this is not the case for simple random sampling, since no marginal stratification has been induced in the simulation procedure.

Stein's LH sampling algorithm, however, does not fully reproduce the correlation in the original SR sample matrix $\mathbf{Y}_R$. The terms in Eq. (2.14) through which correlation can be induced between two LH samples $\mathbf{y}_m^L$ and $\mathbf{y}_{m'}^L$ of any two RVs Y_m and $Y_{m'}$ are the vectors of ranks $\mathbf{r}_m = \text{rank}(\mathbf{y}_m)$ and $\mathbf{r}_{m'} = \text{rank}(\mathbf{y}_{m'})$ of the original SR samples $\mathbf{y}_m$ and $\mathbf{y}_{m'}$ from those two RVs. The corresponding vectors $\mathbf{v}_m$ and $\mathbf{v}_{m'}$ of uniform random numbers in Eq. (2.14), however, are generated independently one from another. For small sample sizes (small S) this can affect the reproduction of a target correlation by the LH sample. Fig. 2.9 provides an assessment of that bias in correlation reproduction for LH sampling from two standard Gaussian RVs with correlation $\rho_{12} = 0.75$ and $\rho_{12} = 0.9$. It can be seen that the mean of the sampling distribution of the correlation coefficient for LH sampling is further away from the target correlation. In addition, the sampling distribution of the correlation coefficient under LH sampling has a similar spread as the corresponding sampling distribution under SR sampling. This is expected, since LH

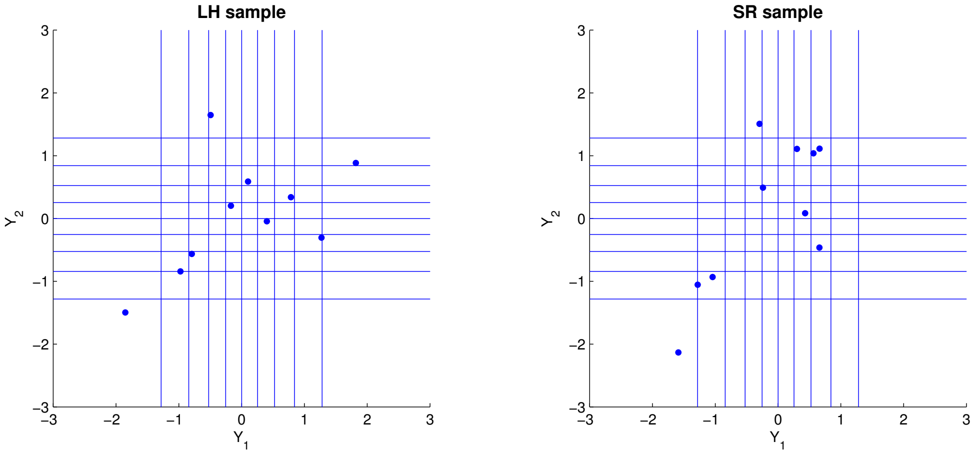


Figure 2.8: Latin hypercube (left) and simple random (right) samples of size $S = 10$ from two standard Gaussian RVs with correlation coefficient $\rho_{12} = 0.75$. Vertical and horizontal lines correspond to deciles of a std Gaussian RV and delineate strata of equal probability.

sampling aims for marginal stratification; hence, there is no expectation for a closer reproduction of that correlation.

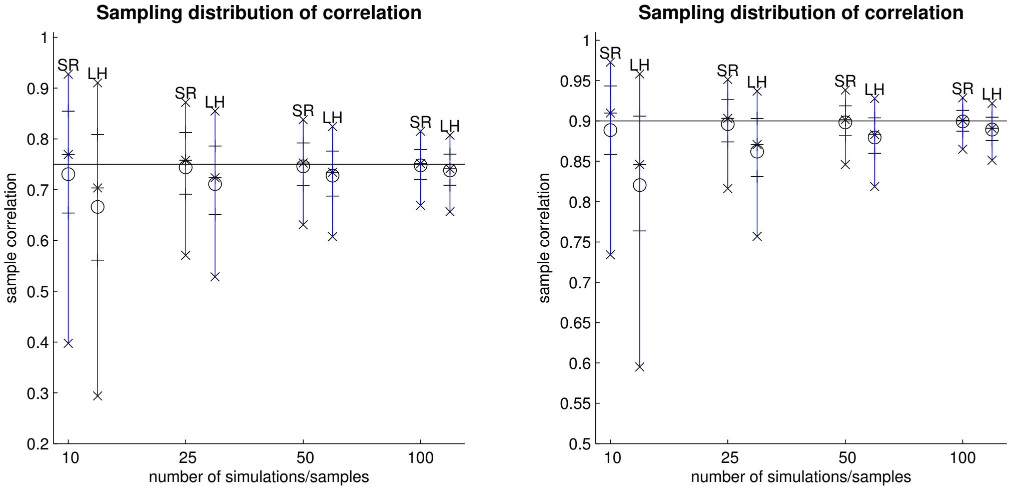


Figure 2.9: Quantiles of the sampling distributions of the correlation coefficient from two standard Gaussian RVs with correlation $\rho_{12} = 0.75$ (left) and $\rho_{12} = 0.9$ (right), obtained via SR and LH sampling using different sample sizes, $S = 10, 25, 50, 100$; see text for details. Horizontal lines running across each plot depict the reference correlations $\rho_{12} = 0.75$ (left) and $\rho_{12} = 0.9$ (right).

The $\mathbf{Y}_L$ can be easily transformed into lognormal $\mathbf{Z}_L$ via Eq.2.13. The same applies for midpoint LH samples.

2.3 Latin hypercube simulation of random fields

Latin hypercube sampling from a multivariate distribution introduced in Section 2.2 is applied here to the spatial analog of a random vector; that is, a random field modeling the spatial distribution of a geo-referenced attribute. More precisely, Subsection 2.3.1 introduces the basic concepts behind a random field and its common parameterization in the multivariate Gaussian case, and Subsection 2.3.2 describes the application of simple random and Latin hypercube sampling in a spatial context for generating realizations from a random field model.

2.3.1 Random field models

Let $y(\mathbf{c})$ denote the value of a geospatial attribute Y, e.g., hydraulic conductivity, at an arbitrary location with coordinate vector $\mathbf{c}$ within a study region A. The (typically unknown) spatial distribution of attribute Y constitutes an attribute surface or field $\{y(\mathbf{c}), \mathbf{c} \in A\}$. In geostatistics, that attribute surface is conceptualized as a realization of a *random field* $\{Y(\mathbf{c}), \mathbf{c} \in A\}$, i.e., an infinite collection of geo-referenced random variables (RVs), one per location (Fig. 2.10); here, $Y(\mathbf{c})$ denotes the RV corresponding to an arbitrary location $\mathbf{c}$, whose realization is denoted as $y(\mathbf{c})$.

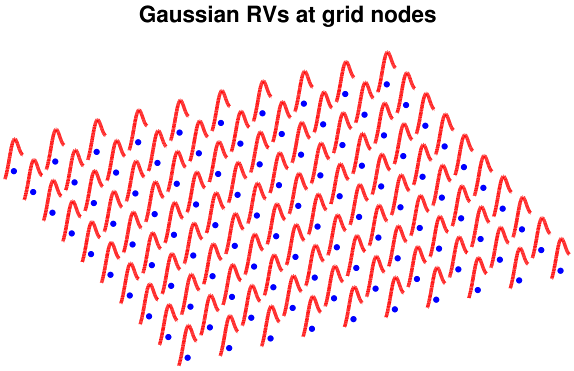


Figure 2.10: Conceptual representation of marginal (local) distributions of geo-referenced random variables comprising (at the limit) a random field; note that the concept of spatial corelation is missing from this figure

A random field is fully characterized by the multivariate distribution of its constituent RVs, also termed the *spatial law* of the random field (Chilès and Delfiner, 1999). The inference of that multivariate distribution in a spatial context, where only one (partial) realization of the random field is available from RVs at sample locations, typically calls for the working hypothesis of second-order stationarity of its (univariate and bivariate) moments.

Under second-order stationarity the expectation of any RV $Y(\mathbf{c})$ is constant within the study area

$$\mathbb{E}\{Y(\mathbf{c})\} = \mu_Y \quad \forall \mathbf{c} \in A \tag{2.15}$$

and the covariance between any two RVs $Y(\mathbf{c})$ and $Y(\mathbf{c}')$ defined at two locations $\mathbf{c}$ and $\mathbf{c}'$ is a function of the magnitude (norm) and possibly orientation (in the anisotropic case) of the separation vector $\mathbf{h} = \mathbf{c} - \mathbf{c}'$ between those two locations

$$\mathbb{C}ov\{Y(\mathbf{c}), Y(\mathbf{c}')\} = \sigma_Y(\mathbf{c} - \mathbf{c}') = \sigma_Y(\mathbf{h}), \;\; \forall \mathbf{c}, \mathbf{c}' \in A \tag{2.16}$$

An arbitrary distance decay function, however, can serve as a valid covariance model if and only if it is positive definite; that is, it satisfies

$$\sum_{m=1}^{M} \sum_{m'=1}^{M} \lambda_m \lambda_{m'} \sigma_Y(\mathbf{c}_m - \mathbf{c}_{m'}) \geq 0 \quad \forall \lambda_m, \lambda_{m'}, \forall \mathbf{c}_m, \mathbf{c}'_m \tag{2.17}$$

ensuring that the variance of a linear combination of RVs $\sum_{m=1}^{M} \lambda_m Y(\mathbf{c}_m)$ is non-negative no matter the weights λ_m and the locations $\mathbf{c}_m$ involved. For this reason, only functions known to be positive definite are used as valid covariance models; see Chilès and Delfiner (1999) for a list of such models.

Parametric distance decay functions $\sigma_Y(\mathbf{h}; \mathbf{p})$ are typically used as covariance models for random fields, parameterized by a sill parameter $\sigma_Y(0)$ – the *a priori* variance σ_Y^2 of the random field Y – and a range parameter r corresponding to the distance at which a zero covariance is attained. A frequently used covariance model is the isotropic exponential covariance function

$$\sigma_Y(h; \sigma_Y^2, r) = \sigma_Y(0) \exp(-3h/r) \tag{2.18}$$

where $h = \|\mathbf{h}\|$ denotes the magnitude (norm) of vector $\mathbf{h}$ and r is the effective range corresponding to 5% of the model sill.

In many cases, including the case study presented at Section 2.5, the spatial variability of the attribute depends on the direction, a situation termed *anisotropy* (Chilès and Delfiner, 1999). A particular case of anisotropic spatial variability is *geometrical anisotropy*, whereby the range of the covariance function changes along different directions while the sill remains the same. In 3D, geometric anisotropy leads to ellipsoidal iso-covariance surfaces, and the locus of 0 covariance values defines an anisotropy ellipsoid; the lengths of the semi-axes of that ellipsoid identify the ranges of the anisotropic covariance model.

When the axes of the anisotropy ellipsoid are aligned with the three cardinal directions and have semi-lengths r_x, r_y and r_z, a unit-sill (with $\sigma_Y(0) = 1$) covariance model becomes a function of the reduced (scaled) separation distance

$$\sigma_Y(\mathbf{h}; r_x, r_y, r_z) = \sigma_Y \left(\sqrt{\left(\frac{h_x}{r_x}\right)^2 + \left(\frac{h_y}{r_y}\right)^2 + \left(\frac{h_z}{r_z}\right)^2} \right) \tag{2.19}$$

instead of the Euclidean distance $h = \sqrt{h_x^2 + h_y^2 + h_z^2}$ used in the isotropic case (Eq 2.18), where h_x, h_y, and h_z are the components of the separation vector $\mathbf{h} = \mathbf{c} - \mathbf{c}'$ along the three cardinal directions.

By defining the component vector $\mathbf{h} = [h_x \; h_y \; h_z]^T$ and arranging the covariance model ranges in a vector $\mathbf{r} = [r_x \; r_y \; r_z]$, Eq. 2.19 can be also expressed in matrix form as

$$\sigma_Y(\mathbf{h}; \mathbf{r}) = \sigma_Y \left(\|\mathbf{R}\mathbf{h}\| \right) \tag{2.20}$$

where $\mathbf{R} = [\text{diag}(\mathbf{r})]^{-1}$ is a (3×3) diagonal matrix with the reciprocal ranges along its diagonal, and $\|\mathbf{Rh}\|$ represents the norm of the scaled version $\mathbf{Rh}$ of the original component separation vector $\mathbf{h}$.

The general case of a anisotropy ellipsoid, however, involves axes that are not aligned with the three cardinal directions. Following (Deutsch and Journel, 1998), we denote as θ_1 the (azimuth) angle of the principal axis from the 0° along the horizontal plane (positive being clock-wise), as θ_2 the (dip) angle from the horizontal plane (positive being downwards), and as θ_3 the plunge angle (increasing counter clock-wise). By arranging these anisotropy orientation parameters in a vector $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \theta_3]$, and denoting as $\mathbf{r} = [r_1 \ r_2 \ r_3]$ the covariance model ranges along the corresponding directions, a general anisotropic covariance model is expressed as

$$\sigma_Y(\mathbf{h}; \mathbf{r}, \boldsymbol{\theta}) = \sigma_Y\left(\|\mathbf{RQh}\|\right) \tag{2.21}$$

where $\mathbf{Q}$ is a (3×3) combined rotation matrix defined as

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_3) & \sin(\theta_3) \\ 0 & -\sin(\theta_3) & \cos(\theta_3) \end{bmatrix} \begin{bmatrix} \cos(-\theta_2) & 0 & \sin(\theta_2) \\ 0 & 1 & 0 \\ -\sin(\theta_2) & 0 & \cos(-\theta_2) \end{bmatrix} \begin{bmatrix} \cos(90-\theta_1) & \sin(90-\theta_1) & 0 \\ -\sin(90-\theta_1) & \cos(90-\theta_1) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.22}$$

with $90-\theta_1$ corresponding to a rotation of the horizontal plane about the Oz axis, leading to new Ox' and Oy' axes, $-\theta_2$ corresponding to a rotation around the Oy' axis, leading to new Ox'' and Oz' axes, and θ_3 corresponding to a rotation around the Ox'' axis leading to the final rotated axes Ox'', Oy'' and Oz'' (Ying, 2001).

An example of such a covariance model is given in Fig. 2.11 using an exponential function with sill $\sigma_Y^2 = 1.79$, and an anisotropy ellipsoid with maximum range $r_1 = 120$ distance units along the 45° direction, $r_2 = 20$ units along the perpendicular direction, and $r_3 = 8$ units along the vertical (hence, $\theta_1 = 45$, and $\theta_2 = \theta_3 = 0$); these parameters are also used in the case study for generating realizations of hydraulic conductivity fields.

In practice, for both isotropic and anisotropic cases, the study area A is typically discretized by a regular grid comprised of $M = M_x M_y$ nodes, where M_x and M_y denote the number of grid nodes along the two cardinal directions. In this case, the continuous random field $\{Y(\mathbf{c}), \mathbf{c} \in A\}$ is approximated by a discrete M-variate spatial random vector $\mathbf{y} = [Y(\mathbf{c}_m), m = 1, \ldots, M]^T$ (Chilès and Delfiner, 1999). In addition, a realization $y(\mathbf{c}_m)$ of the RV $Y(\mathbf{c}_m)$ defined at the m-th grid node $\mathbf{c}_m$ is assumed representative of an elementary area $a_m = a \ \forall m$ around $\mathbf{c}_m$; that is, a represents a constant grid cell of size $|a|$ (Fig. 2.13).

When the random vector $\mathbf{y}$ is assumed multivariate Gaussian, i.e., $\mathbf{y} \sim \mathcal{G}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$, it is fully parameterized by the $(M \times 1)$ expectation vector $\boldsymbol{\mu}_Y = [\mu_Y(\mathbf{c}_m), m = 1, \ldots, M]^T$, and the $(M \times M)$ pairwise covariance matrix $\boldsymbol{\Sigma}_Y = [\sigma_Y(\mathbf{c}_m, \mathbf{c}_{m'}), m = 1, \ldots, M, m' = 1, \ldots, M]$, where $\mu_Y(\mathbf{c}_m) = \mathbb{E}\{Y(\mathbf{c}_m)\}$ and $\sigma_Y(\mathbf{c}_m, \mathbf{c}_{m'}) = \mathbb{C}ov\{Y(\mathbf{c}_m), Y(\mathbf{c}_{m'})\}$. Moreover, under second-order stationarity $\boldsymbol{\mu}_Y$ is a constant vector $\boldsymbol{\mu}_Y = \mu_Y \mathbf{1}$, where $\mathbf{1}$ is

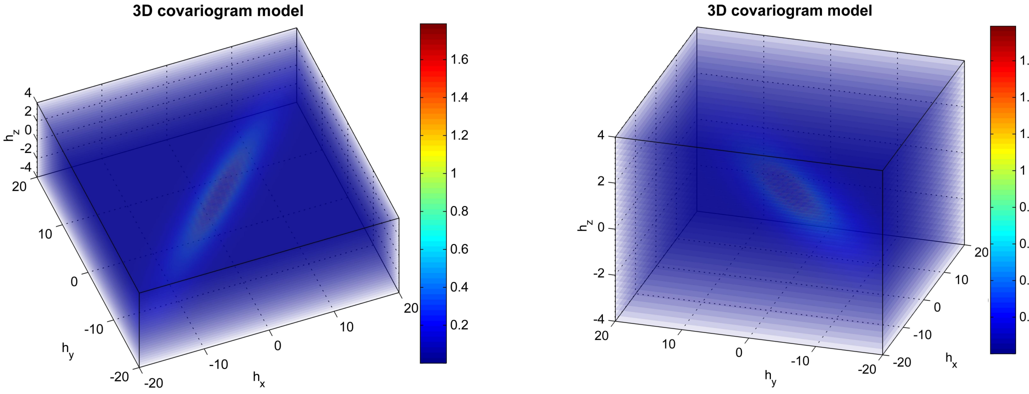


Figure 2.11: Two different viewing perspectives of 3D covariance function grid: any grid node corresponds to a separation vector $\mathbf{h}$ between two locations and is specified by its components $\mathbf{h} = [h_x \; h_y \; h_z]$ along the three cardinal directions. The color scale represents covariance values for each grid node, with the zero separation $\mathbf{h} = [0 \; 0 \; 0]$ placed at the center of the grid.

a $(M \times 1)$ vector of ones, and $\mathbf{\Sigma}_Y$ is constructed using the $M \times M$ distances (in the isotropic case) between the M grid nodes and a valid covariance function $\sigma_Y(h; \boldsymbol{\theta})$, i.e., $\mathbf{\Sigma}_Y = [\sigma_Y(h_{mm'}; \boldsymbol{\theta}), m = 1, \ldots, M, m' = 1, \ldots, M]$, where $h_{mm'} = \|\mathbf{c}_m - \mathbf{c}_{m'}\|$ is the distance between nodes $\mathbf{c}_m$ and $\mathbf{c}_{m'}$. An example of such a covariance matrix $\mathbf{\Sigma}_Y$ is given in Fig. 2.13 using the exponential model of Eq. 2.18 with the parameters used in Fig. 2.12.

2.3.2 LH sampling from random fields

Sampling from a random field amounts to generating S alternative realizations (images in $2D$) of the spatial distribution of a geo-referenced variable Y over the study area A. As in the general multivariate case (Sec. 2.2.2), LH sampling from a random field amounts to (a) generating a spatially correlated SR sample from the random field, and (b) transforming that SR sample into a spatially correlated LH sample with marginally stratified entries.

We denote as $\mathbf{Y}_R = [y_s(\mathbf{c}_m), s = 1, \ldots, S, m = 1, \ldots, M]$ the $(S \times M)$ matrix containing S realizations (a sample of size S) of the random field generated via SR sampling on a set of $M = M_x M_y$ grid nodes discretizing A, where $y_s(\mathbf{c}_m)$ is the s-th simulated value at the m-th location $\mathbf{c}_m$. Alternatively, array $\mathbf{Y}_R$ can be thought of as a collection of M vectors, each containing S simulated values; i.e., $\mathbf{Y}_R = [\mathbf{y}(\mathbf{c}_m), m = 1, \ldots, M]$, where $\mathbf{y}_m = [y_s(\mathbf{c}_m), s = 1, \ldots, S]^T$ denotes a $(S \times 1)$ vector of simulated values at location $\mathbf{c}_m$.

Per Eq. 2.23, the SR sample array $\mathbf{Y}_R$ can be generated from a $(S \times M)$ array of standard Gaussian deviates $\mathbf{W}_R$ and the upper triangular Cholesky factor $\mathbf{L}^T$ of the covariance matrix $\mathbf{\Sigma}_Y$ as

$$\mathbf{Y}_R = \mathbf{W}_R \mathbf{L}^T \tag{2.23}$$

and the results can be displayed as $2D$ images by reshaping arrays $\mathbf{W}_R$ and $\mathbf{Y}_R$ to be of dimensions $(M_x \times M_y \times S)$; see Fig. 2.14.

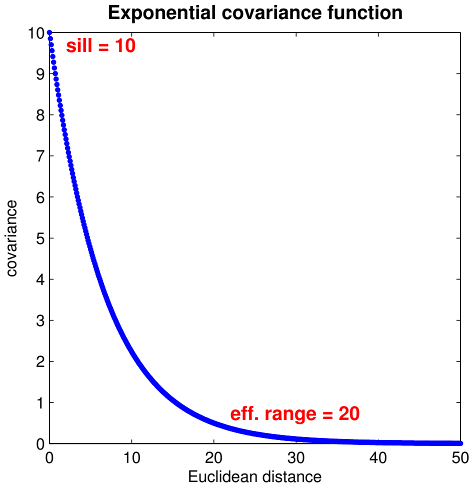


Figure 2.12: Exponential distance decay covariance function with sill and effective range parameters set to 10 and 20, respectively.

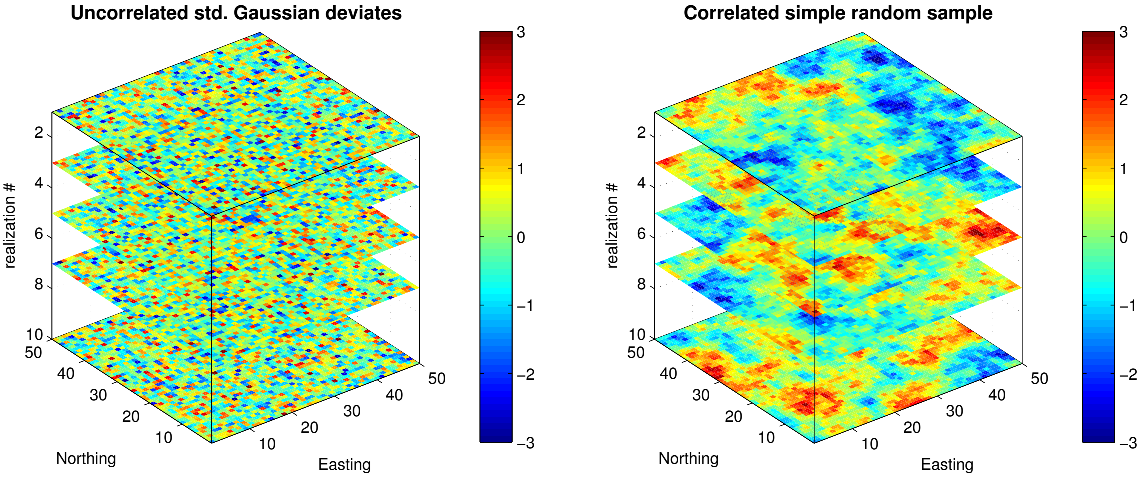


Figure 2.14: Set of 10 realizations (sample of size $S = 10$) of a standard Gaussian random field with an isotropic exponential covariance model with sill 1 and effective range 20 (right), generated from 10 realizations of a, white-noise (with no spatial correlation), standard Gaussian random field (left); see text for details.

Per Eq. 2.14, Stein's method for LH sampling from a random field amounts to transforming the SR sample matrix $\mathbf{Y}_R$ into a $(S \times M)$ LH sample matrix $\mathbf{Y}_L = [y_s^L(\mathbf{c}_m), s = 1, \ldots, S, m = 1, \ldots, S]$ as

$$y_s^L(\mathbf{c}_m) = G^{-1}\left(\frac{r_s(\mathbf{c}_m) - v_s(\mathbf{c}_m)}{S}\right), \quad s = 1, \ldots, S \tag{2.24}$$

where $r_s(\mathbf{c}_m) = \text{rank}(y_s(\mathbf{c}_m))$ denotes the rank of the s-th value $y_s(\mathbf{c}_m)$ of the SR sample $\mathbf{y}(\mathbf{c}_m)$ at location $\mathbf{c}_m$.

In essence, Eq. 2.24 involves a set of S spatially correlated probability values (the argument of the inverse Gaussian CDF), which are also marginally (location-wise) strati-

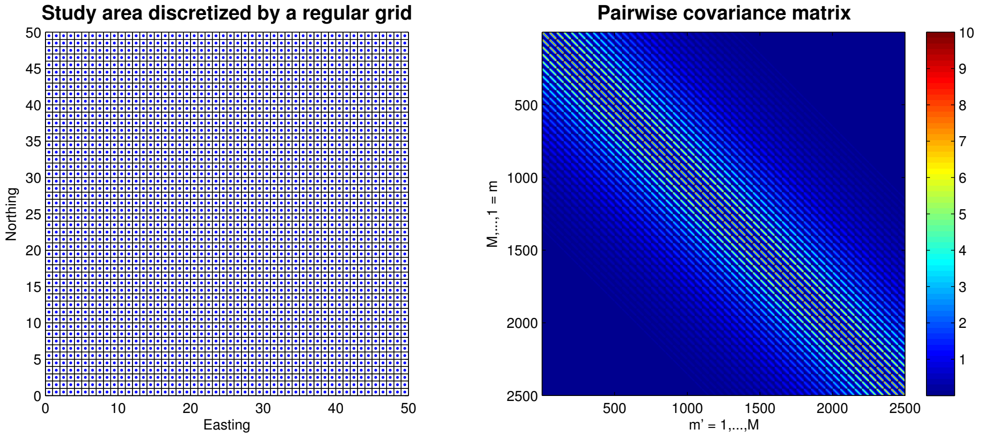
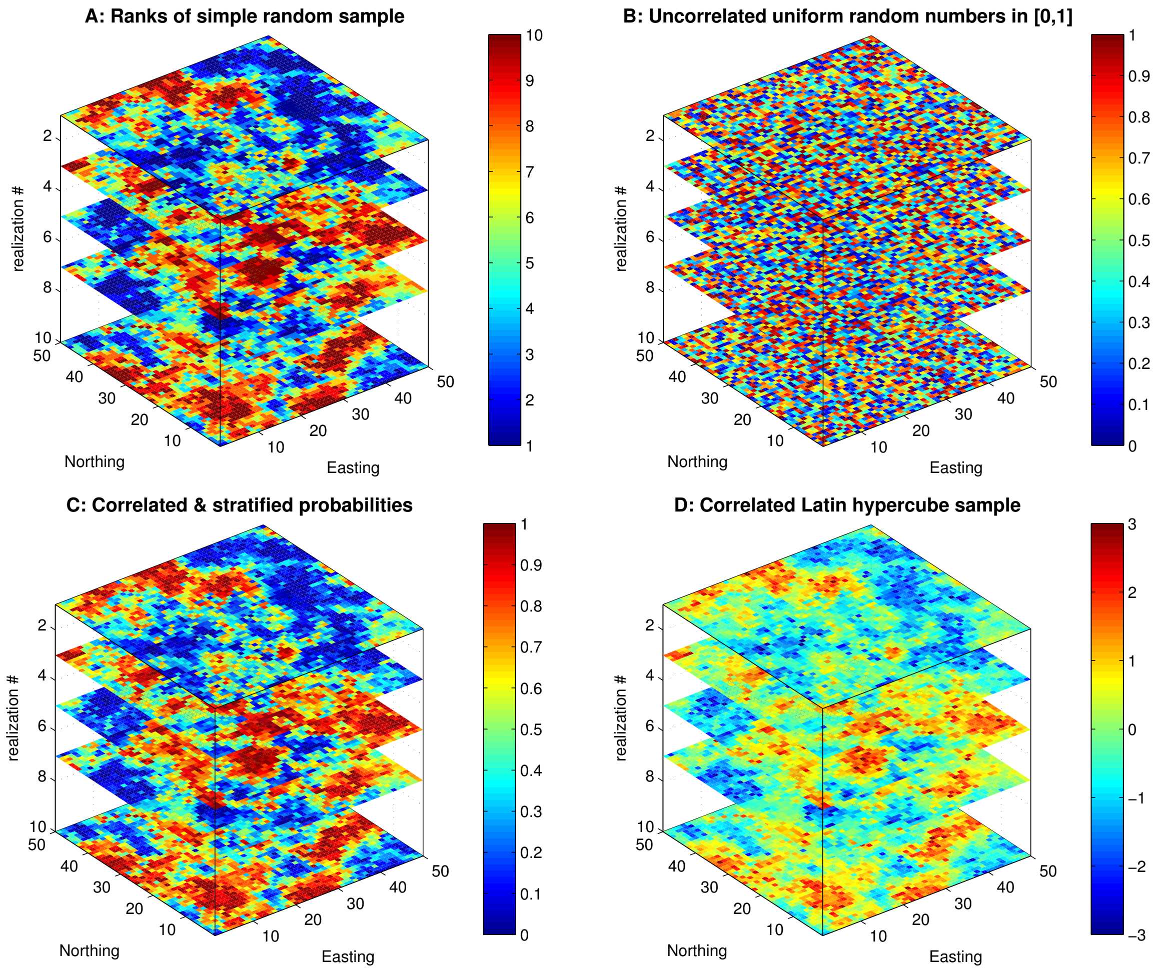


Figure 2.13: **Left:** Example of a 50×50 regular grid of unit spacing. Grid nodes are depicted with blue dots and each such node is the center of a grid cell of size 1×1. The grid origin $(0.5, 0.5)$ is the node at the center of the lower left cell. **Right:** Example of a covariance matrix constructed using distances between all pairs of grid nodes (left) and an isotropic exponential covariance model with sill 10 and effective range 20 (Fig. 2.12).

fied. The rank value $r_s(\mathbf{c}_m)$ identifies the probability stratum associated with an original value $y_s(\mathbf{c}_m)$ of the SR sample at location $\mathbf{c}_m$. The addition of a random number $v_s(\mathbf{c}_m)$ uniformly distributed in $[0, 1]$ furnishes a random probability perturbation within than stratum. Those stratified probability values are then used to derive the corresponding stratified Gaussian quantiles via the inverse CDF; see, Fig. 2.15.



A: Ranks of simple random sample

B: Uncorrelated uniform random numbers in [0,1]

C: Correlated & stratified probabilities

D: Correlated Latin hypercube sample

Figure 2.15: **A:** Rank transformation ($r_s(\mathbf{c}_m)$ in Eq. 2.24) of SR sample of Fig. 2.14, **B:** Uniform random numbers in $[0, 1]$ ($v_s(\mathbf{c}_m)$), simulated independently at each grid node, **C:** Correlated probability values, stratified at each grid node, derived from **A** and **B** ($r_s(\mathbf{c}_m) - v_s(\mathbf{c}_m)$), **D:** Final LH sample of size $S = 10$, whose values are derived as quantiles of a standard Gaussian RV for the stratified probabilities in **C**.

To verify marginal (per grid node) stratification of the LH sample above, Fig. 2.16 shows two scatter plots of simulated values $\mathbf{y}_L(\mathbf{c}_m)$ and $\mathbf{y}_L(\mathbf{c}_m + \mathbf{h})$ extracted from the LH sample of Fig. 2.15D at two randomly selected pairs of nodes separated by distances $h = \|\mathbf{h}\|$ set to 1 and 4. The covariance values for those two scatter plots approach their model counterparts $\sigma_Y(1)$ and $\sigma_Y(4)$ given by the unit-sill exponential covariance function with effective range 20. In addition, LH sampling yields marginally stratified simulated values at any grid node, as expected by theory.

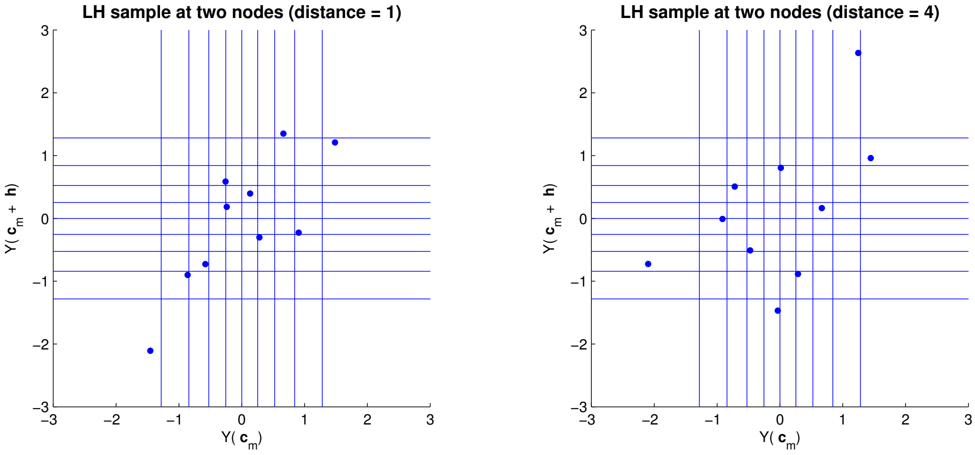


Figure 2.16: Two scatter plots of simulated values extracted from the LH sample of Fig. 2.15D at two pairs of randomly selected nodes separated by distance 1 (left) and 4 (right). Vertical and horizontal lines correspond to deciles of a std Gaussian RV and delineate strata of equal probability; see text for details.

2.4 Synthetic hydrogeological case studies

In this Section two simple hydrogeological case studies are adopted - one two dimensional isotropic and one three dimensional anisotropic - comparing simple random and Latin hypercube sampling from a lognormal random field. In both case studies lognormal random fields model the spatial distribution of saturated hydraulic conductivity in hydrogeological context involving flow and transport in a heterogeneous porous medium. Regarding both case studies respectively, in Subsections 2.4.1 and 2.5.1, the two sampling methods are evaluated in terms of their ability to reproduce model conductivity field parameters. In Subsections 2.4.2 and 2.5.2, simple random and Latin hypercube are evaluated here again in terms of their ability to reproduce ensemble concentration statistics computed by solving the hydrogeological model of flow and transport for the above large set of hydraulic conductivity fields.

2.4.1 2D hydraulic conductivity

A two-dimensional synthetic groundwater flow system is considered in this case study, similar to that used in Zhang and Pinder (2003). The dimensions of the flow system are 5100 m by 5100 m discretized into a 51 × 51 grid with uniform rectangular cells of size 100 m by 100 m (Fig. 2.17). Porosity was assumed constant throughout the domain and equal to 0.25, and the parametrization of hydraulic conductivity is described below. Flow boundary conditions consisted of constant head of 0 m at the four corner cells and a constant head of 250 m at the central cell of the domain (Fig. 2.17). No flow conditions ($\partial h/\partial n = 0$) were assigned to the rest of the domain boundaries, whereas the Modflow code (McDonald and Harbaugh, 1988) was used in this study to obtain the steady state flow solution.

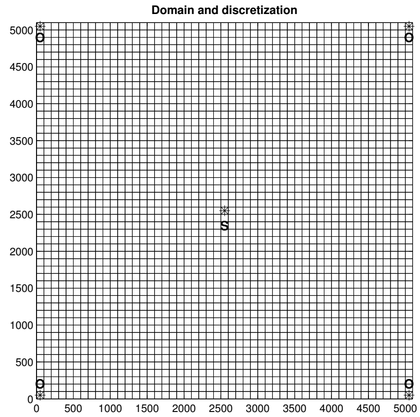


Figure 2.17: Two-dimensional simulation domain of size 5100 m by 5100 m discretized into a 51×51 grid with uniform cell size of 100 m by 100 m.

A second-order stationary and isotropic lognormal random field is adopted with parameters borrowed from Sudicky et al. (2010). More precisely, the mean and variance of log conductivity are taken as $\mu_Y = -5.64$ and $\sigma_Y^2 = 1.79$, respectively, corresponding to conductivity statistics $\mu_Z = 0.0087$ m/sec and $\sigma_Z^2 = 0.0194^2$ (m/sec)^2. The semivariogram of log conductivity is assumed to be of exponential form, with no nugget effect, and effective range 1000 m, corresponding to one fifth of the domain extent along the cardinal directions. Two realizations of this random field model at the nodes of the grid shown in Fig. 2.17 are given in Fig. 2.18.

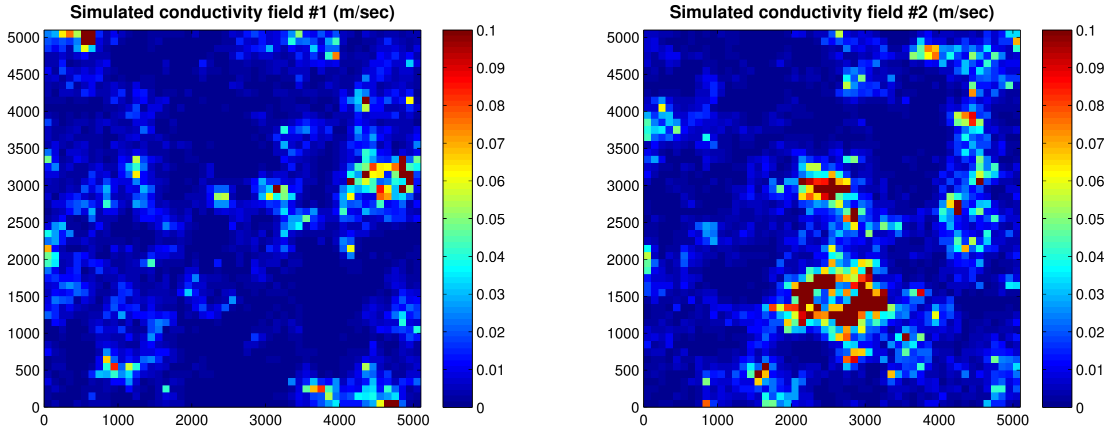


Figure 2.18: Two realizations of a lognormal random field; see text for details.

Reference ensemble statistics consist of: (i) the ensemble average field – a constant equal to $\mu_Z = 0.0087$, (ii) the ensemble standard deviation field – a constant equal to $\sigma_Z = 0.0194$, (iii) the ensemble pairwise correlation between grid cells – this involves entries of the upper triangular part of the covariance matrix $\mathbf{\Sigma}_Z$ (iv) the ensemble global proportion of grid nodes with conductivity values exceeding a threshold of $z = 0.007$

m/sec corresponding to the 0.695 quantile of a lognormal distribution with mean $m_Z = 0.0087$ and variance $\sigma_Z^2 = 0.0194^2$ – hence a constant $p = 1 - 0.695 = 0.305$, and (v) the ensemble local probability field of exceeding that conductivity threshold of $z = 0.007$ m/sec.

Three sampling or simulation methods are considered in this case study for generating realizations of a lognormal hydraulic conductivity field with the parameterization (mean, variance and correlation length) given above on the 51×51 simulation grid. These methods include simple random (SR) sampling, Latin hypercube (LH) sampling and its midpoint version (LHM). In terms of sample size or number of realizations per method, five such sizes are considered; namely, $S = 10, 25, 50, 75,$ and 100. Once a sample, say of size $S = 25$, is generated, the discrepancy between the statistics of the simulated ensemble and the reference conductivity statistics listed above is quantified using the root mean squared error (RMSE) for summary statistics (mean, variance, correlation).The computation of such error statistics is repeated over a set of $I = 500$ batches of realizations, with each batch containing the same sample size, $S = 25$ for example, thus estimating the sampling distributions of RMSE values for each sample size and for each method; these distributions are presented in terms of their means and medians, as well as their 75% and 95% probability intervals. The better the reproduction of a reference statistic from simulations of a sampling method with a given sample size, the narrower the sampling distribution of the resulting, say, RMSE values, and the smaller (closer to 0) the center of that distribution.

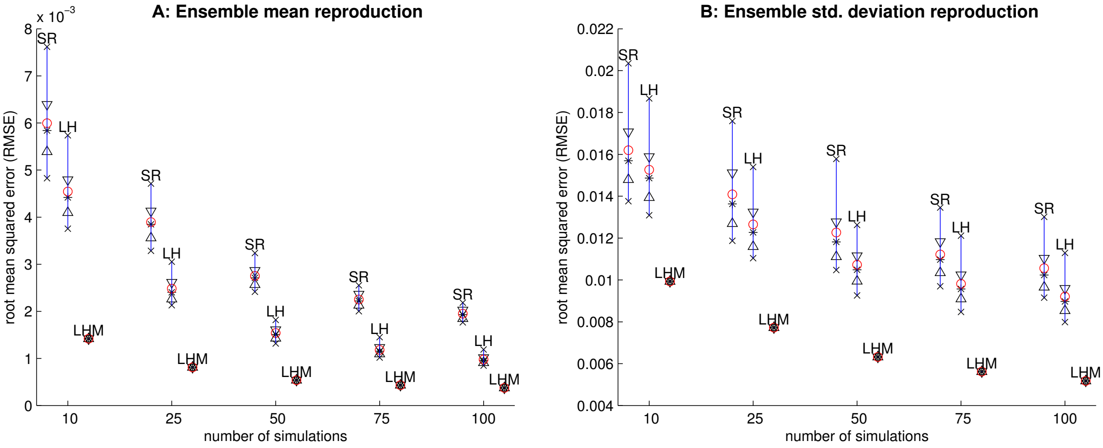


Figure 2.19: Reproduction of reference ensemble mean (A) and ensemble standard deviation (B) hydraulic conductivity fields from various sampling methods; see text for details.

More precisely, the reproduction of the reference ensemble mean and ensemble standard deviation of hydraulic conductivity from the three sampling methods and the five sample sizes considered is shown in Fig. 2.19. Reproduction is quantified in terms of the sampling distribution of RMSE between reference and simulated ensemble conductivity statistics over the 51×51 grid cells. In this and all subsequent figures, 75% RMSE or MAE depending on the statistic selected) probability intervals are depicted with horizontal line segments, whereas 95% probability intervals with × symbols; median RMSE values are depicted as asterisks ($*$), whereas mean values as circles ($\circ$). From Fig. 2.19, it can be readily appreciated that midpoint LH sampling (LHM) yields the closest reproduction for both the ensemble mean and standard deviation statistics, followed by standard LH sampling. This is expected, since LH sampling aims at marginal stratification, hence should best reproduce marginal hydraulic conductivity statistics at each grid node.

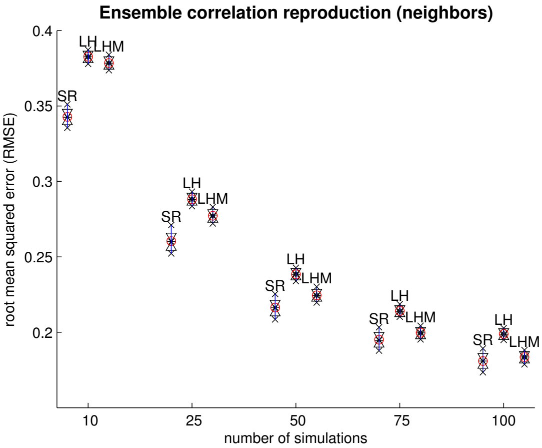


Figure 2.20: Reproduction of reference hydraulic conductivity ensemble correlation matrix between first-order neighboring grid cells.

Figure 2.20 shows the reproduction of ensemble correlation for first-order neighboring grid nodes for the three sampling methods and the five sample sizes considered. First-order neighbors of an arbitrary grid node are here defined using 8-point connectivity; that is, such neighbors include the four immediate neighbors in the four cardinal directions, as well as the four grid nodes to the main diagonals of the central node under consideration. Reproduction is quantified in terms of the sampling distribution of RMSE computed between the pairwise correlations of the reference and simulated hydraulic conductivity fields at first-order neighboring grid cells. From Fig. 2.20, it can be easily appreciated that LH and LHM sampling exhibit a bias in that short-scale correlation reproduction. It should also be noted that all three methods yield the same RMSE reproduction when the correlation for both reference and simulated conductivity fields is computed for all possible pairs of grid nodes; i.e., when evaluating the reproduction of the entire upper diagonal part of the conductivity covariance matrix $\mathbf{\Sigma}_Z$.

Figure 2.21A shows the reproduction of the global proportion of grid nodes with conductivity values greater than 0.007 m/sec for the three sampling methods and the five sample sizes considered. This conductivity threshold corresponds to the 0.695 quantile of a lognormal distribution with mean $m_Z = 0.0087$ and variance $\sigma_Z^2 = 0.0194^2$ – hence a constant $p = 1 - 0.695 = 0.305$. Reproduction is quantified here in terms of the sampling distribution of RMSE computed between the reference and simulated global proportions of conductivity values exceeding that threshold. LH sampling yields the best reproduction for all sample sizes in this case, whereas LHM demonstrates a bias of not efficiently

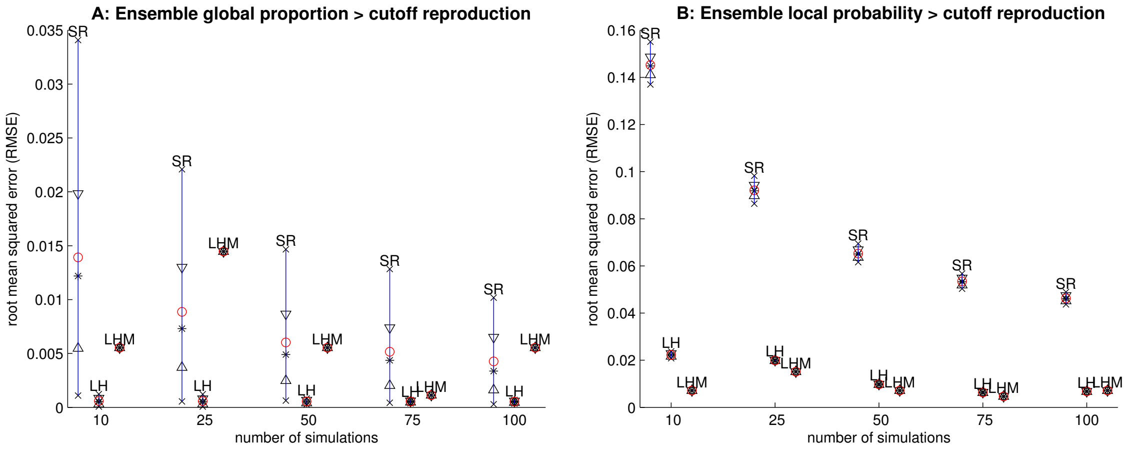


Figure 2.21: Reproduction of reference global (A) and local (B) ensemble probability of exceeding a conductivity threshold value of 0.007 m/sec, derived for various sampling methods.

reproducing RMSE values probably because it does not sample the extreme values of the distribution. Figure 2.21B shows the reproduction of that probability of exceedance $p = 0.305$ at each simulation grid node. Reproduction is quantified here in terms of the sampling distribution of RMSE computed between the reference and simulated local proportions of conductivity values exceeding that threshold; the RMSE here is calculated via a double summation over nodes and realizations. As expected, LH and LHM sampling yield the best reproduction in this case, due to the marginal stratification they impose on the simulated conductivity realizations at each node. Contrary to the global case, LHM here shows no bias in the results since the high correlation of nodes accounted for the calculation of local probability of exceedance shows small relevance to the non inclusion of extreme values of the distribution.

2.4.2 2D solute concentration

For the solute transport problem, an initial concentration equal to 0 is assumed throughout the model domain. At time $t = 0$, a contaminant is introduced at the central cell, along the upstream constant head boundary, with constant concentration $C_0 = 100$ mg/l (Fig. 2.17). No transport conditions ($\partial C/\partial n = 0$) are assigned along the domain boundaries. Longitudinal and transverse horizontal dispersivities are assumed to be equal to 5 m and 0.5 m, respectively. In terms of software, the MT3D code (Zheng, 1990) was used to obtain breakthrough curves at the four observation wells located at the four corner nodes of the domain, as well as the solute transport solution up to time $t = 2 \cdot 10^6$ sec. Reference ensemble statistics for solute concentration are derived from a set of 10000 solutions of the transport problem based on the 10000 hydraulic conductivity realizations generated in the previous subsection via SR sampling. Two such concentration realizations derived from the conductivity realizations shown in Fig. 2.18 are given in Fig. 2.22.

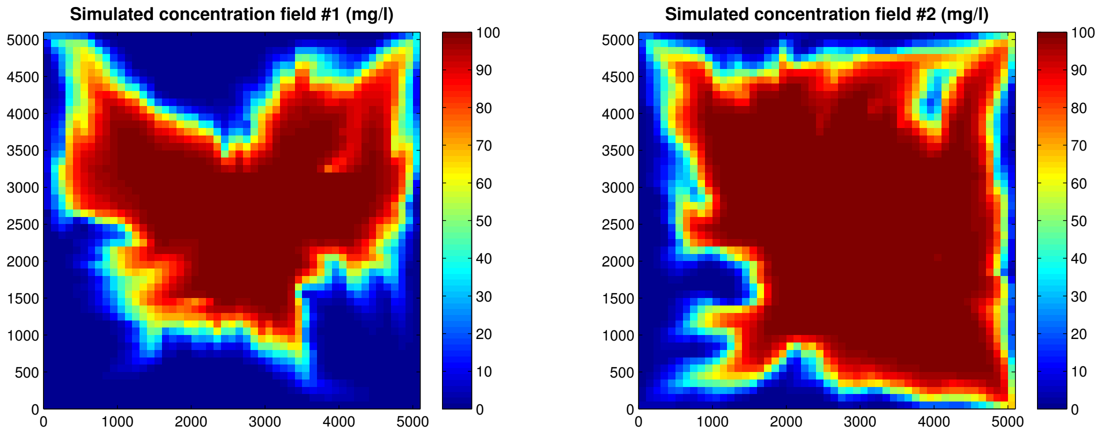


Figure 2.22: Two solute concentration realizations corresponding to the two hydraulic conductivity realizations shown in Fig. 2.18; see text for details.

The performance of the various sampling methods considered in this work for sample sizes $S = 10, 25, 50, 75, 100$ is then quantified in terms of reproduction of these reference ensemble concentration statistics via the sampling distribution of error summary statistics (RMSE and/or MAE). Reference statistics for solute concentration consist of: (i) the ensemble average field shown in Fig. 2.23 upper left, (ii) the ensemble standard deviation field also shown in Fig. 2.23 upper left, (iii) the ensemble pairwise correlation between grid nodes – this involves all the entries of the upper triangular part of the concentration correlation matrix without focusing on short-scale reproduction due to the non-stationarity of the concentration field (that is, each covariance is computed separately for each absolute separation vector and compared with the relative reference value), and (iv) the ensemble distribution of first arrival times (Fig. 2.23 bottom) at any of the four observation wells (Fig. 2.17).

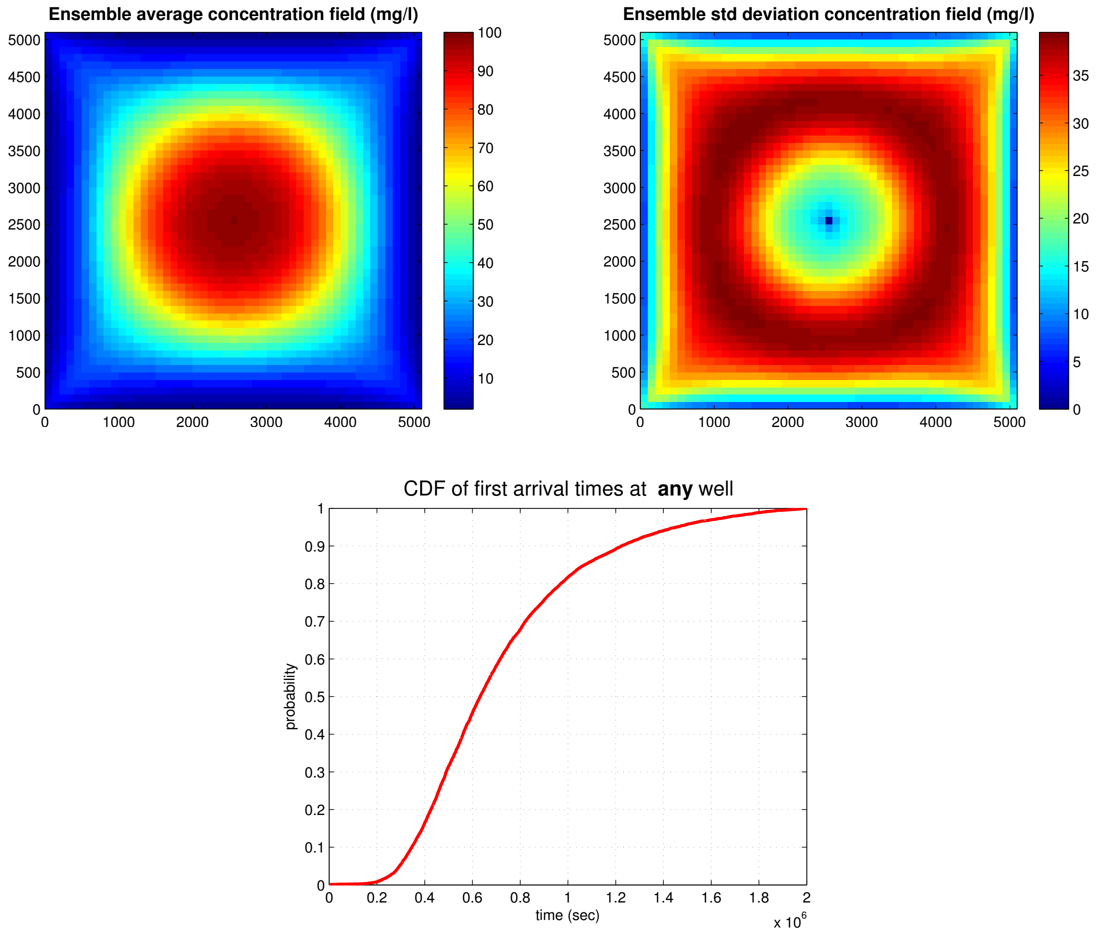


Figure 2.23: Reference ensemble average (top left), ensemble standard deviation (top right) concentration fields, and ensemble first arrival time CDF (bottom), computed from 10000 concentration realizations derived from 10000 hydraulic conductivity realizations generated via simple random sampling; see text for details.

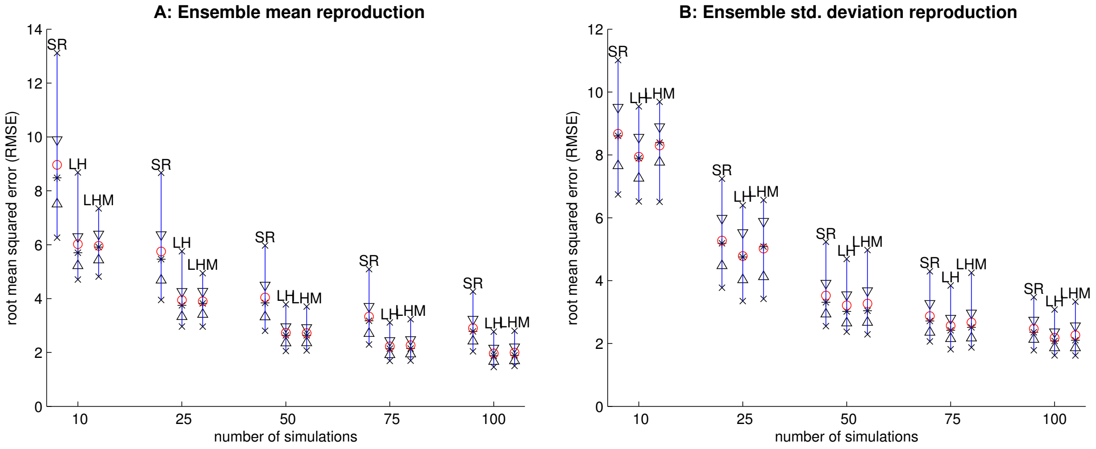


Figure 2.24: Reproduction of reference ensemble mean (A) and ensemble standard deviation (B) concentration fields (shown in Fig. 2.23) from various sampling methods; see text for details.

Figure 2.24 shows the reproduction of the ensemble average and ensemble standard deviation concentration fields (Fig. 2.23) for the three sampling methods and the five sample sizes considered. Reproduction is quantified here in terms of the sampling distribution of RMSE between reference and simulated ensemble concentration statistics over the 51×51 grid cells. It is easily appreciated that LH and LHM yield a better reproduction of these ensemble fields than SR sampling, particularly in the case of the ensemble average concentration field.

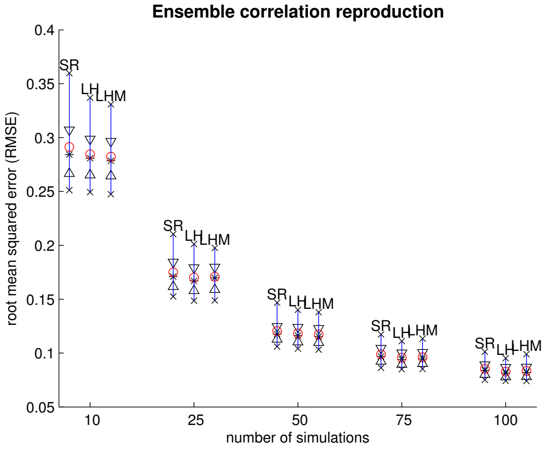


Figure 2.25: Reproduction of reference concentration ensemble correlation matrix.

Figure 2.25 gives the reproduction of the reference concentration correlation matrix (upper triangular part) for the three sampling methods and the given sample sizes. Reproduction is quantified here in terms of the sampling distribution of RMSE computed between the pairwise correlations of the reference and simulated concentration fields at all pairs of grid cells. LH and LHM sampling achieve a similarly better reproduction than SR sampling.

Last, Fig. 2.26 gives the reproduction of the reference cumulative distribution function (CDF) of first arrival times at any of the four observation wells at the corners of the domain, and the corresponding CDFs obtained from the three sampling methods and the five sample sizes considered. Discrepancy between any two CDFs is quantified in terms of MAE between their corresponding percentiles, and the sampling distribution of this statistic is established over $I = 500$ batches of realizations as before.

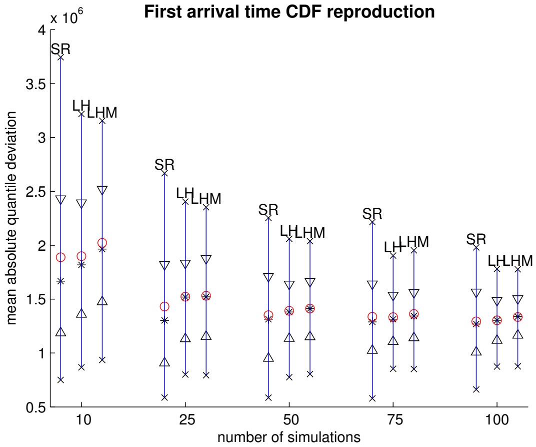


Figure 2.26: Reproduction of reference first arrival time cumulative distribution function (CDF), derived for various sampling methods; see text for details.

2.5 Anisotropic three dimensional hydrogeological case study

Latin Hypercube is here again compared with Simple Random sampling in generating realizations from a lognormal random field modeling the spatial distribution of saturated hydraulic conductivity in an anisotropic 3D hydrogeological context. The generated conductivity fields are further used to derive solute concentration fields through flow and transport simulation. The two sampling methods are then compared again in terms of their ability to reproduce ensemble statistics of the solute concentration field. Midpoint Latin hypercube was not examined hereafter because the excessively targeted sampling of values, concluding in the exclusion of any randomness, was considered sub-optimal and beyond the Monte Carlo framework of this dissertation. On this regard, all the results of Sec. 2.4.1 show that midpoint Latin hypercube hold almost zero variance in the reproduction of RMSE values, thus minimizing the inherent randomness of the simulation procedure.

The evaluation of the performance of the two sampling methods in generating representative realizations from a (log)Gaussian random field is performed using a three-dimensional synthetic hydrogeological problem involving flow and transport in a porous medium. The dimensions of the synthetic flow system are 820 m by 820 m by 180 discretized into a $41 \times 41 \times 9$ grid with uniform uniform cells of size 20 m by 20 m by 20 m (Fig. 2.27).

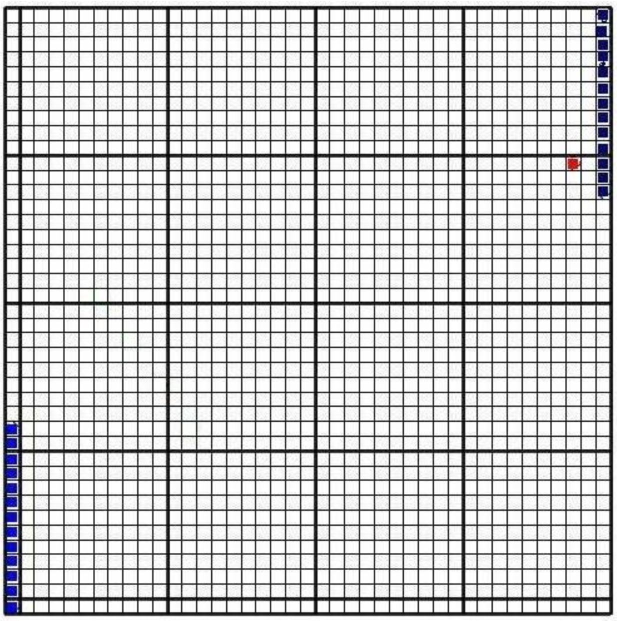


Figure 2.27: Top view of the three-dimensional simulation domain of size 820 m by 820 m by 180 discretized into a $41 \times 41 \times 9$ grid with uniform cells of size 20 m by 20 m by 20 m.

The flow boundary conditions consist of constant head of 0 m at the cells $B(x, y, z)$, where $x = 1$, $1 \leq y \leq 13$, $1 \leq z \leq 9$ (lower part of the left hand boundary), and a constant head of 10 m at the cells $B(x, y, z)$, where $x = 41$, $29 \leq y \leq 41$, $1 \leq z \leq 9$ (upper part of the right hand boundary). No flow conditions ($\partial h / \partial n = 0$) are assigned to the rest of the domain boundaries. Porosity is assumed constant throughout the domain and equal to 0.25. The Modflow code (McDonald and Harbaugh, 1988) is used in this study to obtain the steady state flow solution.

2.5.1 3D hydraulic conductivity

A second-order stationary, lognormal anisotropic random field is adopted with parameters borrowed from (Sudicky et al., 2010). More precisely, the mean and variance of log conductivity are taken as $\mu_Y = -5.64$ and $\sigma_Y^2 = 1.79$, respectively, corresponding to conductivity statistics $\mu_Z = 0.0087$ m/sec and $\sigma_Z^2 = 0.0194^2$ (m/sec)^2. The semivariogram of log conductivity is assumed to be of exponential form, with no nugget effect. The effective range is 450 m along the major axis oriented 45° increasing clockwise from the zero azimuth (North - South), 80 m along the minor axis perpendicular to the major axis, and 35 m along the z axis. This implies a six fold anisotropy along the horizontal plane. No further anisotropy and/or directionality is considered along the vertical z axis where the effective range corresponds to the one fifth of the vertical extend of the domain. Two realizations of this random field model are given in Fig. 2.28.

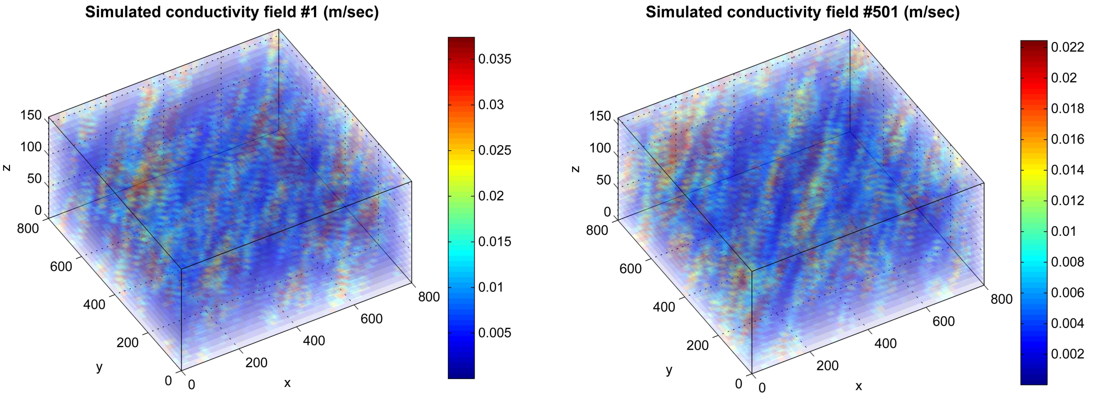


Figure 2.28: Two realizations of a lognormal random field; see text for details.

Reference ensemble statistics are derived from a set of 1000 realizations of hydraulic conductivity generated via simple random (SR) sampling, and consist of: (i) the ensemble average field, and (ii) the ensemble standard deviation field. Both ensemble statistics are depicted in Fig. 2.29.

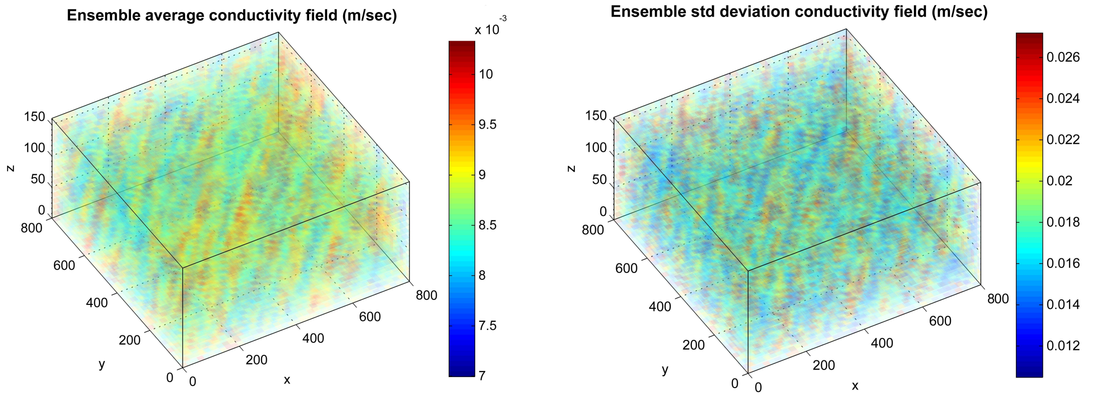


Figure 2.29: Reference ensemble average (left) and ensemble standard deviation (right) conductivity fields, computed from 1000 hydraulic conductivity realizations generated via simple random sampling; see text for details.

Simple random (SR) and Latin hypercube (LH) are considered in this case study for generating realizations of a lognormal hydraulic conductivity field with the parameterization (mean, variance and correlation length) given above on the $41 \times 41 \times 9$ simulation grid. In terms of sample size or number of realizations per method, three such sizes are considered; namely, $S = 20, 50$, and 80. Once a sample, say of size $S = 20$, is generated, the discrepancy between the statistics of the simulated ensemble and the reference conductivity statistics listed above is quantified using the root mean squared error (RMSE) for the summary statistics (i)mean, and (ii)variance.

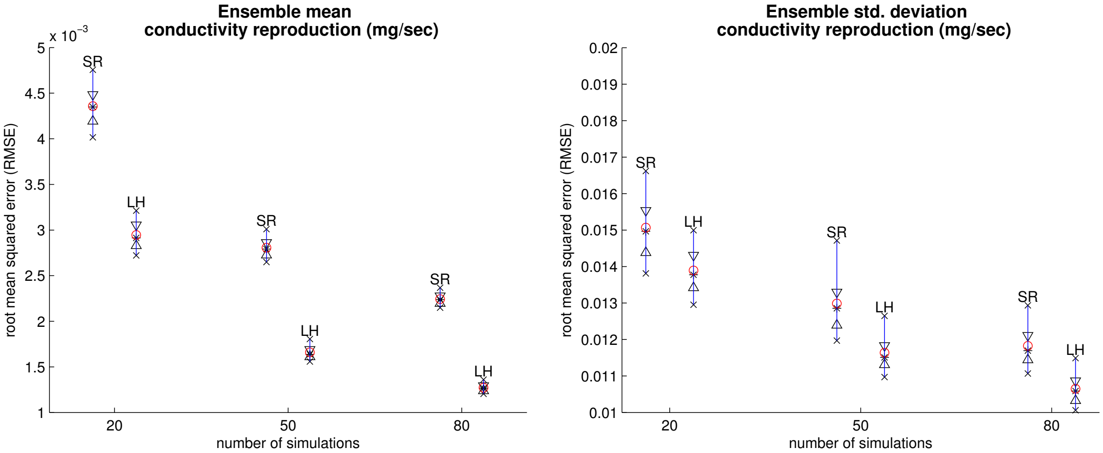


Figure 2.30: Reproduction of reference ensemble mean (A) and ensemble standard deviation (B) hydraulic conductivity fields from various sampling methods; see text for details.

More precisely, the reproduction of the reference ensemble mean and ensemble standard deviation of hydraulic conductivity from the two sampling methods and the three sample sizes considered is shown in Fig. 2.30. Reproduction is quantified in terms of the sampling distribution of RMSE between reference and simulated ensemble conductivity statistics over the $41 \times 41 \times 9$ grid cells.

From Fig. 2.30, it can be readily appreciated that LH sampling (LH) yields the closest reproduction for both the ensemble mean and standard deviation statistics. This is expected, since LH sampling aims at marginal stratification, hence should best reproduce marginal hydraulic conductivity statistics at each grid node.

2.5.2 3D Solute concentration

For the solute transport problem, an initial concentration equal to 0 is assumed throughout the model domain. At time $t = 0$, a contaminant is introduced at the injection well $C(x, y, z)$, depicted as a red dot in Fig. 2.27, where $x = 38, y = 30, z = 5$ that is located near the top right corner of the 3D domain, close to the upstream constant head boundary, with constant concentration $C_0 = 1000$ mg/l. Longitudinal, transverse horizontal and transverse vertical dispersivities are assumed to be equal to 5 m and 0.5 m and 0.05 m respectively.

Reference ensemble statistics for solute concentration are derived from a set of 1000 solutions of the transport problem based on the 1000 hydraulic conductivity realizations

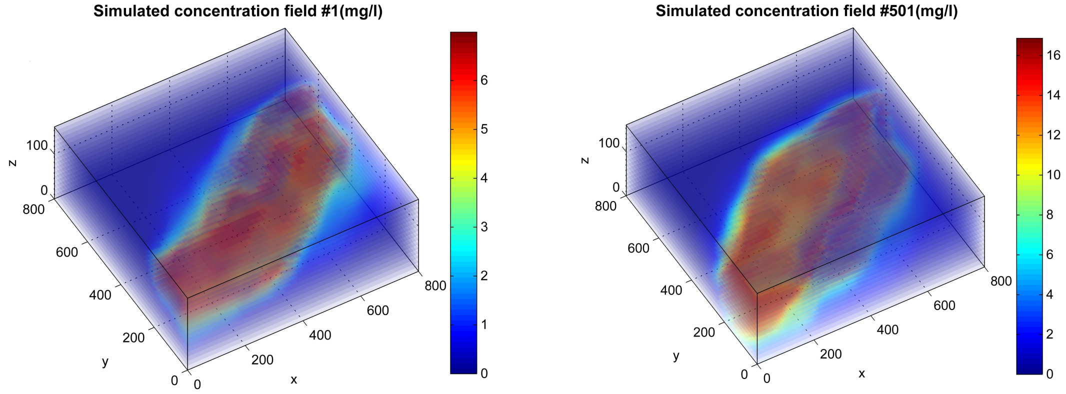


Figure 2.31: Two solute concentration realizations corresponding to the two hydraulic conductivity realizations shown in Fig. 2.28; see text for details.

generated in the previous subsection via SR sampling. Two such concentration realizations derived from the conductivity realizations shown in Fig. 2.28 are given in Fig. 2.31.

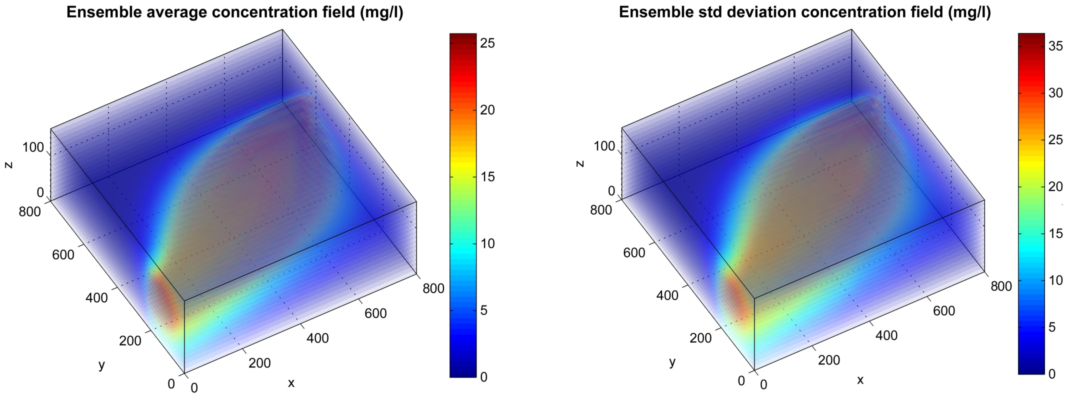


Figure 2.32: Reference ensemble average (A) and ensemble standard deviation (B) concentration fields, computed from 1000 concentration realizations derived from 1000 hydraulic conductivity realizations generated via simple random sampling; see text for details.

The performance of the various sampling methods considered in this work for sample sizes $S = 20, 50, 80$ is then quantified in terms of reproduction of these reference ensemble concentration statistics via the sampling distribution of error summary statistics (RMSE). Reference statistics for solute concentration consist of: (i) the ensemble average field shown in Fig. 2.32 A, (ii) the ensemble standard deviation field also shown in Fig. 2.32 B.

Figure 2.33 shows the reproduction of the ensemble average and ensemble standard deviation concentration fields (Fig. 2.32) for the two sampling methods and the three sample sizes considered. Reproduction is quantified here in terms of the sampling distribution of RMSE between reference and simulated ensemble concentration statistics over

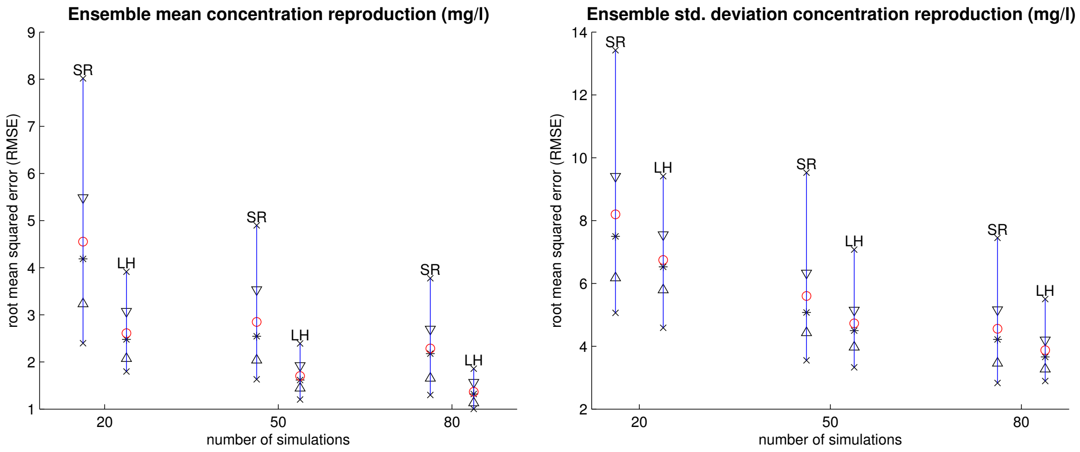


Figure 2.33: Reproduction of reference ensemble mean (A) and ensemble standard deviation (B) concentration fields (shown in Fig. 2.32) from various sampling methods; see text for details.

the 41 × 41 × 9 grid cells. It is easily appreciated that LH yield a better reproduction of these ensemble fields than SR sampling, particularly in the case of the ensemble average concentration field.

2.6 Conclusions and Discussion

The spatial distribution of saturated hydraulic conductivity is often parameterized in terms of a lognormal random field model in stochastic two or three dimensional hydrogeological case studies. Simulated realizations of such a random field are often used along with physically-based models of flow and transport in a Monte Carlo framework for evaluating the uncertainty in the spatial distribution of solute concentration due to the uncertainty in the spatial distribution of hydraulic conductivity. In practical applications, however, classical Monte Carlo simulation, or simple random (SR) sampling, from a random field model can quickly become computationally expensive. A more computationally efficient alternative to classical Monte Carlo simulation is Latin hypercube (LH) sampling, aiming at reducing the number of realizations that are too similar by chance. In a spatial context, LH sampling aims at generating representative realizations from a random field model, where term representative implies realizations spanning efficiently the range of possible attribute realizations corresponding to the multivariate probability distribution (lognormal in this work) associated with the random field.

This Chapter is an introduction to Monte Carlo simulation procedure and moreover to Stein's LH method (Stein, 1987), both in a spatial context. The purpose here is to get the reader familiar with the concept of stratified simulation by applying Stein's method and highlighting its benefits compared to classical Monte Carlo. Apart from the simple introductory character, this chapter introduces a synthetic hydrogeological case study. The case study includes the computation of arrival times of the pollutant at certain observation wells. It is a rather more stringent evaluation of the proposed method, which might reveal the small added value of the method due to the high degree of non-linearity.

Moreover, it is natural for the reader to expect a discussion regarding conditioning to specific data (like the measurements data at the four corners of the domain for both case studies); this is treated separately in Chap. 5.

More specifically, the performance of simple random (SR) and Latin hypercube (LH) sampling is compared in two case studies (one 2D and an anisotropic 3D) within a hydrogeological context involving flow and transport in a heterogeneous porous medium. The two methods are evaluated as per their ability of reproducing reference conductivity ensemble statistics. The statistics considered for the first case study included the ensemble mean, standard deviation, global and local probabilities of threshold exceedance, as well as short-scale correlation, whereas for the second case study the ensemble mean and standard deviation. The reproduction of the same statistics for the ensemble concentration field resulting from solving a flow and transport boundary value problem for each hydraulic conductivity realization, was also evaluated in the second part of the two case studies respectively.

As expected for the case of hydraulic conductivity, LH sampling (and its Midpoint extension regarding the 2D case study) achieved the best reproduction of marginal reference statistics, for the ensemble mean and standard deviation (Fig. 2.19 and Fig. 2.30). Regarding the results of the 2D case alone, LH sampling achieved a better reproduction of the marginal reference local probability of exceeding a relatively high concentration threshold, close to the 70-th percentile of the global distribution (Fig. 2.21B). Note that the above reductions in sampling variability brought by LH sampling as compared to SR sampling might be even more visible and appreciable had the error analyses been conducted using Gaussian realizations instead of their lognormal counterparts.

In terms of solute concentration, LH sampling (and its Midpoint extension regarding for the 2D case study) showed significantly better reproduction of the ensemble mean and standard deviation concentration fields than SR sampling, in the case of the ensemble average concentration field (Fig. 2.24 and Fig. 2.33). Regarding the two dimensional case studies, the three sampling methods considered yielded comparable reproductions of the ensemble concentration correlation (Fig. 2.25) and lastly, LH sampling showed the best reproduction of the reference distribution of first arrival times at any of the four observation wells at the four corner nodes of the simulation grid (Fig. 2.26).

It is easily appreciated considering the results for both two dimensional and anisotropic three dimensional case studies, that LH sampling yields a better reproduction of these ensemble average fields than SR sampling, for both model inputs and outputs. It could be further argued that LH sampling offers smaller sampling variability for the same number of simulated realizations than SR sampling. Consequently, LH sampling could lead to more efficient uncertainty propagation with fewer model runs due to more representative inputs, thus reducing the time and computer resources for such an endeavor.

Chapter 3

Stratified Sampling from Random Fields

In this Chapter two alterations of an innovative sampling method are introduced, scarcely known in the geostatistical literature, here named "Stratified Likelihood sampling" (SL) and "Minimum Energy sampling" (ME), in which representative realizations are generated by exploring in a systematic way the structure of the multivariate distribution function itself. *Term representative implies realizations spanning efficiently the range of possible attribute realizations corresponding to the multivariate probability distribution (lognormal in this work) associated with the random field.* All three controlled or stratified sampling methods (LH, SL and ME), are capable of generating representative realizations from (log)Gaussian random fields, i.e., spanning efficiently the range of values corresponding to the (log)Gaussian multivariate probability distribution. Although such realizations often serve as parameters for physical process simulators, existing controlled sampling methods do not account for model sensitivity; hence, they need not yield representative realizations of model outputs. To address this shortcoming, controlled sampling methods are embedded within a two-step simulation procedure. The first step involves stratified sampling at a set of control points where attribute values are expected to exert a large impact on model predictions and/or where uncertainty in such predictions is expected to be largest. In the second step, control point samples are used to generate attribute realizations over the entire study region using classical geostatistical simulation. The proposed controlled, two-step geostatistical simulation approach is further evaluated via two synthetic case studies involving physically-based simulation of flow and transport in a porous medium with known boundary and initial conditions over a simple geometrical domain; their performance is evaluated for different sample sizes (number of realizations) in terms of the reproduction of ensemble statistics of hydraulic conductivity and solute concentration computed from a very large ensemble set generated via simple random sampling. The results show that Latin hypercube, stratified likelihood and minimum energy sampling are more efficient than simple random sampling, in that they can overall reproduce to a similar extent statistics of the conductivity and concentration fields, yet with smaller sampling variability than the latter.

The remainder of the chapter is structured as follows: Section 3.1 introduces stratified likelihood sampling, Section 3.3 describes minimum energy sampling and presents a simple hydrogeological case study of flow and transport in a heterogeneous porous medium highlighting the benefits of the proposed method, and Section 3.2 illustrates the two-step simulation procedure and compares the controlled simulation methods via a hydrogeological case study also underlying the need to account for the uncertainty related to respective Hydrogeological flow and transport model applied.

3.1 Stratified Likelihood Sampling

An alternative method for representative sampling from a multivariate (log)normal distribution, namely, stratified likelihood (SL) sampling, is described hereafter. First proposed in a conference proceedings paper by Switzer (2000) in the context of Gaussian random fields, brought a different perspective on the generation of representative realizations from a multivariate Gaussian probability distribution. Kyriakidis and Gaganis (2013) enhanced the method in a spatial context further investigating its performance in a hydrogeological case study. The basic idea is to generate in a structured (controlled or stratified) way representative realizations from the multi-Gaussian distribution by capitalizing on the geometrical properties of the associated likelihood function; hence, the name stratified likelihood sampling coined in this paper. The metric of representativity of a realization is taken here as the Mahalanobis distance of that realization to the multivariate (multi-point) attribute expectation, conditional or not to sample data. Unlike LH sampling, SL sampling does not aim at marginal stratification but at generating representative realizations by exploring in a systematic way the underlying multi-Gaussian likelihood model. The basic idea is to select via stratification representative points (attribute realizations) in a hyper-dimensional space from that likelihood. Although the method is in principle also applicable to non-Gaussian distributions, the focus of this paper is primarily on the multivariate (log)normal case which is frequently encountered in hydrogeological applications. For a more comprehensive approach to efficient sampling that capitalizes on approximate knowledge regarding the links between model inputs and model outputs for a wide range of applications, the reader is referred to Scheidt and Caers (2009), and Caers (2011).

In what follows, Subsection 3.1.1 introduces the basic concepts behind multi-Gaussian likelihood sampling, and Subsection 3.1.2 describes the extension of the multivariate lognormal stratified likelihood sampling.

3.1.1 Multi-Gaussian likelihood sampling

At the heart of SL sampling lies the squared Mahalanobis distance d_M^2 between a realization $\mathbf{y}$ and the expectation vector $\boldsymbol{\mu}_Y$, defined as Johnson and Wichern (1998)

$$d_M^2(\mathbf{y}) = [\mathbf{y} - \boldsymbol{\mu}_Y]^T \boldsymbol{\Sigma}_Y^{-1} [\mathbf{y} - \boldsymbol{\mu}_Y] = d_M^2(\mathbf{y}, \boldsymbol{\mu}_Y; \boldsymbol{\Sigma}_Y) \tag{3.1}$$

where $\boldsymbol{\Sigma}_Y^{-1}$ denotes the inverse of the covariance matrix $\boldsymbol{\Sigma}_Y$, and the notation $d_M^2(\mathbf{y}, \boldsymbol{\mu}_Y; \boldsymbol{\Sigma}_Y)$ could be used to explicate that the distance between $\mathbf{y}$ and $\boldsymbol{\mu}_Y$ is parameterized in terms of $\boldsymbol{\Sigma}_Y$. In a nutshell, the Mahalanobis distance between two points $\mathbf{y}$ and $\boldsymbol{\mu}_Y$ in a M-dimensional scatter plot takes into account the geometry of the variability of that scatter plot as encapsulated by the covariance matrix $\boldsymbol{\Sigma}_Y$. More specifically, the squared Mahalanobis distance is a version of a standardized squared Euclidean distance, whose components are measured along the directions specified by the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}_Y$. The component of distance along the m-th direction specified by the m-th eigenvector of $\boldsymbol{\Sigma}_Y$ quantifies distance between the m-th entries of $\mathbf{y}$ and $\boldsymbol{\mu}_Y$ in terms of standard deviation along that direction given by the m-th eigenvalue of the $\boldsymbol{\Sigma}_Y$. Although the Mahalanobis distance is not tied to any multivariate distribution, it is often used in a multi-Gaussian context, due to the fact that a multivariate Gaussian distribution is fully specified by the expectation vector $\boldsymbol{\mu}_Y$ and the covariance matrix $\boldsymbol{\Sigma}_Y$.

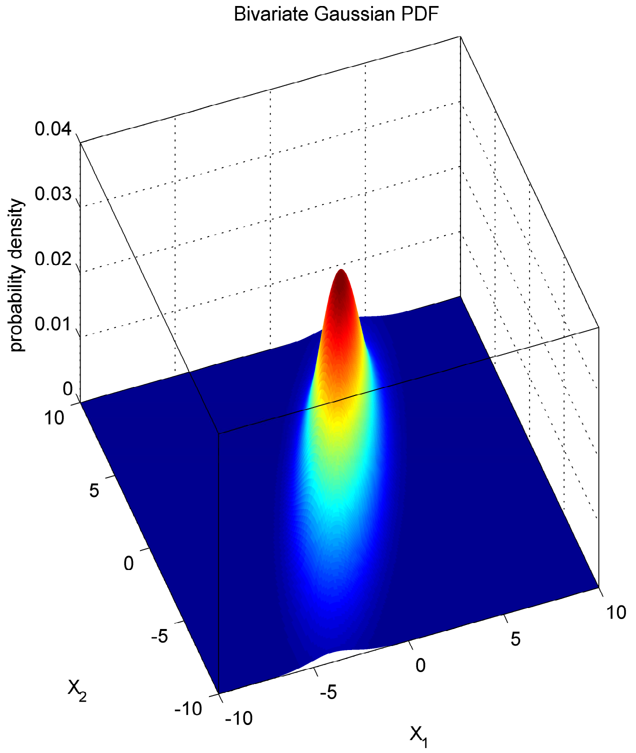


Figure 3.1: Bivariate Gaussian density function; it should be noted that for 3 variables the iso-likelihood contours become ellipsoids

For the case of a M-variate Gaussian PDF $\mathcal{G}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$, (bivariate example illustrated in Fig. 3.1) the squared Mahalanobis distance $d_M^2(\mathbf{y})$ determines the likelihood of realization $\mathbf{y}$ up to a normalizing constant, since the multivariate Gaussian PDF is a function of $d_M^2(\mathbf{y})$ Johnson and Wichern (1998). Two realizations $\mathbf{y}$ and $\mathbf{y}'$ with the same likelihood also correspond to the same squared Mahalanobis distance $d_M^2(\mathbf{y})$. Such realizations can be represented as points lying on a hyper-ellipsoid (bivariate example illustrated in Fig. 3.2) centered at $\boldsymbol{\mu}_Y$ with axes orientations identified by the M eigenvectors of the covariance matrix $\boldsymbol{\Sigma}_Y$, and axes lengths proportional to the square root of the corresponding M eigenvalues of $\boldsymbol{\Sigma}_Y$; the proportionality constant is $d_M(\mathbf{y})$. For a more recent and comprehensive exposition of the links between the Mahalanobis distance and multi-Gaussian likelihood, the reader is referred to (Bourgault, 2012).

In addition, for a M-variate Gaussian PDF the squared Mahalanobis distance d_M^2

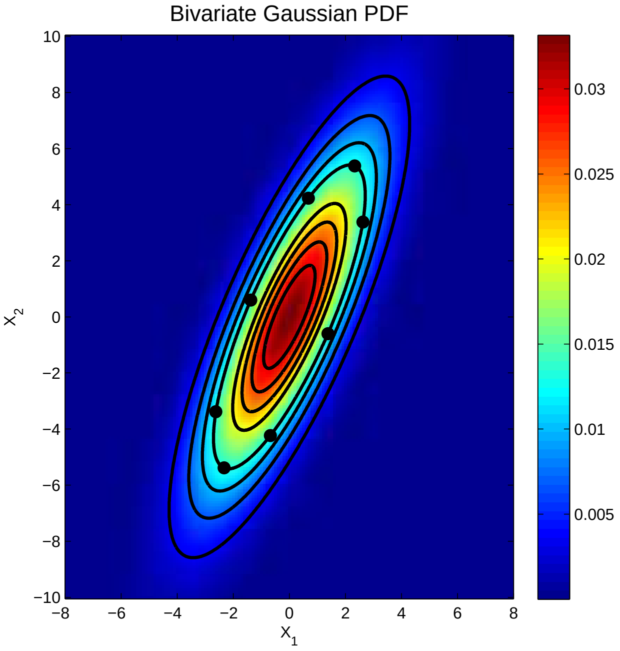


Figure 3.2: Point corresponding to equally likely pairs of realizations on a bivariate Gaussian density function

follows a chi square distribution with M degrees of freedom $d_M^2 \sim \chi^2(M)$, and the associated Mahalanobis distance d_M follows a chi distribution with the same degrees of freedom $d_M \sim \chi(M)$. This implies that selecting, for example, the 0.95 quantile $x_{0.95}$ of a chi distribution with M degrees of freedom, amounts to selecting a hyper-ellipsoid of constant likelihood; Fig. 3.3. There are multiple realizations $\mathbf{y}$ on this hyper-ellipsoid indexed by $x_{0.95}$, all of which are equally likely. SL sampling amounts to generating efficiently such multivariate Gaussian realizations by selecting representative points on likelihood-indexed hyper-ellipsoids.

More precisely, it can be shown that a realization $\mathbf{y}$ from a multivariate Gaussian PDF $\mathcal{G}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ can be represented in terms of a chi deviate x and a unit norm realization $\tilde{\mathbf{u}}$ or point on the surface of a unit hyper-sphere, as Switzer (2000)

$$\mathbf{y} = \mathbf{C}^T \mathbf{w} + \boldsymbol{\mu}_Y = \mathbf{C}^T x \tilde{\mathbf{u}} + \boldsymbol{\mu}_Y \tag{3.2}$$

where $\mathbf{C}$ is the (upper triangular) Cholesky factor of the covariance matrix $\boldsymbol{\Sigma}_Y$, $\mathbf{w}$ is a $(M \times 1)$ vector of standard Gaussian deviates, $\tilde{\mathbf{u}}$ is a $(M \times 1)$ vector of uniform deviates in $[-1, 1]$ constrained to a unit norm, $\|\tilde{\mathbf{u}}\| = \sqrt{\tilde{\mathbf{u}}^T \tilde{\mathbf{u}}} = 1$, and x is a scalar from a chi distribution with M degrees of freedom, $x \sim \chi(M)$, independent of the entries of $\tilde{\mathbf{u}}$. Term $x\tilde{\mathbf{u}}$ in Eq. (3.2) is essentially a generalization of the polar method for generating a realization $\mathbf{w}$ of M uncorrelated standard Gaussian deviates Kroese et al. (2011): the entries of $\tilde{\mathbf{u}}$ define a random point on the surface of a unit hyper-sphere in M dimensions, or equivalently a random orientation, and the chi deviate x defines a random radius; Fig 3.4.

Vector $\tilde{\mathbf{u}}$ can be efficiently generated by constraining a $(M \times 1)$ vector $\mathbf{v}$ of uncorrelated

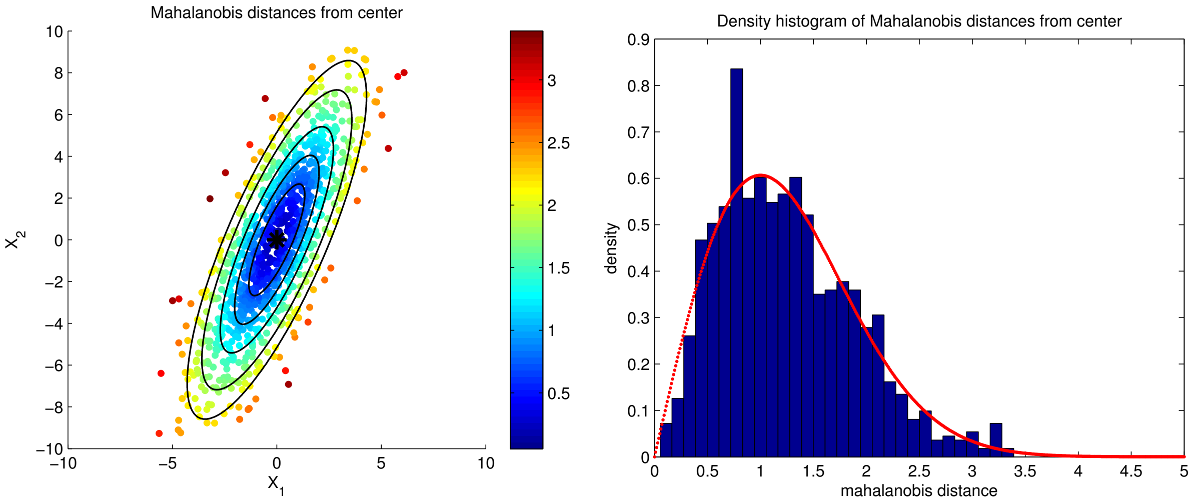


Figure 3.3: Mahalanobis distances from center $\boldsymbol{\mu}_Y$ (left) and the relevant density histogram of Mahalanobis distances following a chi distribution with M degrees of freedom (right)

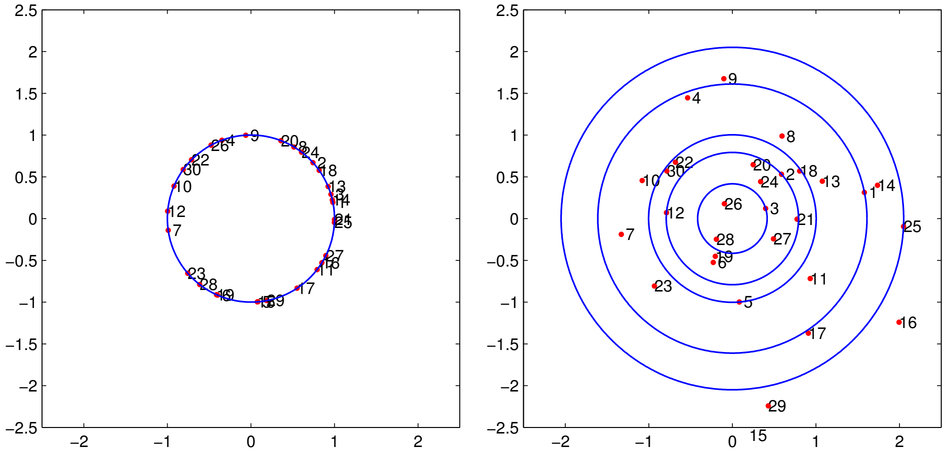


Figure 3.4: Simulation of random points on the surface of a unit hyper-sphere in M dimensions (circle on the left), simulation of S random orientations from the center and displacement of points from unit hyper-sphere to new hyper-spheres of random radii (right)

standard Gaussian deviates to have a unit norm, as $\tilde{\mathbf{u}} = \mathbf{v}/\sqrt{\mathbf{v}^T\mathbf{v}}$ Kroese et al. (2011). Multiplication by the Cholesky factor $\mathbf{C}$ in Eq. (3.2) amounts to rotating and scaling the entries of $\tilde{\mathbf{u}}$ so that they now define a point lying on the surface of a M-dimensional hyper-ellipsoid whose geometry is specified by the eigenvectors and eigenvalues of the covariance matrix $\mathbf{\Sigma}_Y$.

Note that scalar x serves as an index of likelihood-based hyper-ellipsoid selection; this likelihood is linked to the Mahalanobis distance d_M associated with a given hyper-ellipsoid. Indeed, the procedure of simulating a chi deviate x and using Eq. (3.2) to generate a (zero mean for simplicity) realization $\mathbf{y}$ from a multivariate Gaussian PDF $\mathcal{G}(\mathbf{0}, \mathbf{\Sigma}_Y)$ ensures that realization $\mathbf{y}$ has a Mahalanobis distance equal to the chi square deviate x^2, since

$$\mathbf{y}^T\mathbf{\Sigma}_Y^{-1}\mathbf{y} = \tilde{\mathbf{u}}^T\mathbf{C}x(\mathbf{C}^T\mathbf{C})^{-1}x\mathbf{C}^T\tilde{\mathbf{u}} = x^2\tilde{\mathbf{u}}^T\tilde{\mathbf{u}} = x^2 = d_M^2(\mathbf{y})$$

using the fact that $(\mathbf{C}^T\mathbf{C})^{-1} = \mathbf{C}^{-1}(\mathbf{C}^T)^{-1}$, and $\tilde{\mathbf{u}}$ has unit norm.

Equation (3.2) can be generalized to the simulation of a $(S \times M)$ matrix $\mathbf{Y}$ with a simple random (SR) sample of size S from a multivariate Gaussian PDF $\mathcal{G}(\boldsymbol{\mu}_Y, \mathbf{\Sigma}_Y)$, as

$$\mathbf{Y} = \mathbf{W}\mathbf{C} + \mathbf{M}_Y = \mathit{diag}(\mathbf{x})\tilde{\mathbf{U}}\mathbf{C} + \mathbf{M}_Y \tag{3.3}$$

where $\mathit{diag}(\mathbf{x})$ is a $(S \times S)$ diagonal matrix having as diagonal entries a set of S deviates $\mathbf{x} = [x_n, n = 1, \dots, N]^T$ from a chi distribution with M degrees of freedom. Note that the s-th row of matrix $\tilde{\mathbf{U}}\mathbf{C}$ is multiplied by the s-th chi deviate x_n; that is, the (s, s) entry of matrix $\mathit{diag}(\mathbf{x})$ or the s-th element of $\mathbf{x}$. Term $\tilde{\mathbf{U}}$ is a $(S \times M)$ matrix with uniform deviates in $[-1, 1]$, all rows of which are constrained to have a unit norm; that is, if $\tilde{\mathbf{u}}_s$ denotes the s-th row of $\tilde{\mathbf{U}}$, then $\sqrt{\tilde{\mathbf{u}}_s^T\tilde{\mathbf{u}}_s} = 1$. Matrix $\tilde{\mathbf{U}}$ can be generated as vector $\tilde{\mathbf{u}}$ above by: (i) simulating a $(S \times M)$ matrix $\mathbf{V}$ of uncorrelated standard Gaussian deviates, (ii) forming the $(S \times 1)$ vector $\|\mathbf{V}\|$ containing the norms of its S rows, and (iii) computing the product $[\mathit{diag}(\|\mathbf{V}\|)]^{-1}\mathbf{V}$, with $[\mathit{diag}(\|\mathbf{V}\|)]^{-1}$ denoting a $(S \times S)$ diagonal matrix with the reciprocals of the entries of vector $\|\mathbf{V}\|$ along its diagonal. Term $\mathit{diag}(\mathbf{x})\tilde{\mathbf{U}}$ is a $(S \times M)$ matrix $\mathbf{W}$ with uncorrelated standard Gaussian deviates, which can be transformed as $\mathbf{W}\mathbf{C}$ into a SR sample $\mathbf{Y}$ of size S from a M-variate Gaussian PDF $\mathcal{G}(\boldsymbol{\mu}_Y, \mathbf{\Sigma}_Y)$. Matrix $\mathbf{Y}$ can be finally transformed into a SR sample $\mathbf{Z}$ from a M-variate lognormal distribution as $\mathbf{Z} = \exp(\mathbf{Y})$.

Equation (3.3) is essentially an alternative but equivalent version of Eq. (2.23), now expressed in terms of S chi deviates (entries of vector $\mathbf{x}$) and S sets of M uniform deviates in $[-1, 1]$ (rows of matrix $\tilde{\mathbf{U}}$) representing points constrained to lie on a S-dimensional unit hyper-sphere. Considering S points uniformly distributed on the surface of a M-dimensional unit hyper-sphere (rows of $\tilde{\mathbf{U}}$) and rotating/stretching them as $\tilde{\mathbf{U}}\mathbf{C}$ to lie onto the surface of a M-dimensional hyper-ellipsoid, however, leads to a non-uniform point distribution on the hyper-ellipsoid (clumping of points near the poles) due to the non-constant curvature of the latter (Borovkov, 1994; Kroese et al., 2011). One has to keep in mind, however, that it is the matrix $\mathbf{W}$ of uncorrelated standard Gaussian deviates in Eq. (3.3) that is generated as $\mathit{diag}(\mathbf{x})\tilde{\mathbf{U}}$. That matrix $\mathbf{W}$ is then transformed into a matrix $\mathbf{Y}$ of correlated Gaussian deviates via $\mathbf{W}\mathbf{C}$, a procedure well-known to produce realizations from a multivariate Gaussian distribution. Viewed alternatively, since the chi distribution approaches a Gaussian one as M increases Kendall (1945), the S chi deviates in vector $\mathbf{x}$ are more likely to attain values near the center of that distribution. This implies that uniformly distributed points on the unit hyper-sphere (rows of $\tilde{\mathbf{U}}$) are less likely to be displaced (due to scaling by the entries of $\mathbf{x}$) onto hyper-spheres with much larger radii or with locus far away from the center of the unit hyper-sphere; hence, differential clumping near the poles of a hyper-ellipsoid due to subsequent rotation of those points by $diag(\mathbf{x})\tilde{\mathbf{U}}\mathbf{C}$ affects fewer points.

3.1.2 Multivariate lognormal stratified likelihood sampling

The objective here is to generate a representative sample of size S from a multivariate lognormal PDF $\mathcal{L}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$, which is tantamount to generating a representative sample from a multivariate Gaussian PDF $\mathcal{G}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ per the monotonicity of the transformation $Z = \exp(Y)$. A closer inspection of Eq. (3.3) reveals that the terms through which such an objective could be achieved are: (i) the matrix $\tilde{\mathbf{U}}$ of S sets of uniform deviates defining S points on the M-dimensional unit hyper-sphere, generated as $\tilde{\mathbf{U}} = diag(\|\mathbf{V}\|)^{-1}\mathbf{V}$, where $\mathbf{V}$ is a matrix of uncorrelated standard Gaussian deviates and $\|\mathbf{V}\|$ contains the S scalar norms of the S rows of $\mathbf{V}$, and (ii) the vector $\mathbf{x}$ of S chi-square deviates displacing these S points on hyper-spheres with different radii. Stratified likelihood (SL) sampling amounts to: (i) generating a representative matrix $\tilde{\mathbf{U}}_L$ whose S rows define S points that span efficiently the M-dimensional unit hyper-sphere, and (ii) generating a representative (Latin hypercube) sample $\mathbf{x}_L$ of size S from a chi distribution with M degrees of freedom. A trivariate example of a random and a stratified point design of 100 samples, along with the relevant cdf of the nearest neighbor distances are illustrated in Fig. 3.5. It is easily comprehended that the stratified point design has the maximum possible nearest neighbor distances among all sampled values.

Task (ii) can be accomplished via a straightforward application of Eq. (2.5) for LH sampling from a univariate (chi in this case) distribution. For task (i), (Switzer, 2000) proposed to induce stratification across the rows of $\tilde{\mathbf{U}}$ via stratification of the entries of each column of $\mathbf{V}$; this is the approach followed in this work, too. Since the entries of any column of $\mathbf{V}$ are generated independently of the entries of any other column, (Switzer, 2000) proposed to employ Eq. (2.5) for univariate LH sampling M times to generate a $(S \times M)$ matrix $\mathbf{V}_L$ whose columns contain S stratified standard Gaussian deviates. It should be noted, here, that Switzer's proposal for addressing task (i) does not guarantee that the S points defined by the rows of matrix $\tilde{\mathbf{U}}_L$ span efficiently, e.g., repel each other hence cover uniformly, the unit hyper-sphere. Consequently, the current preliminary implementation of SL sampling offered in this work is approximate and therefore sub-optimal. Section 3.3 present an improvement on this front by stratifying across the rows of $\tilde{\mathbf{V}}$ via a multiple steps procedure maximizing the nearest neighbor distances of the generated samples (state of convergence).

Simulation flowchart: In summary, stratified likelihood (SL) sampling, as implemented in this work, for generating a sample of size S from a M-variate lognormal PDF $\mathcal{L}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$; that is, S representative realizations from that PDF, proceeds in the following steps:

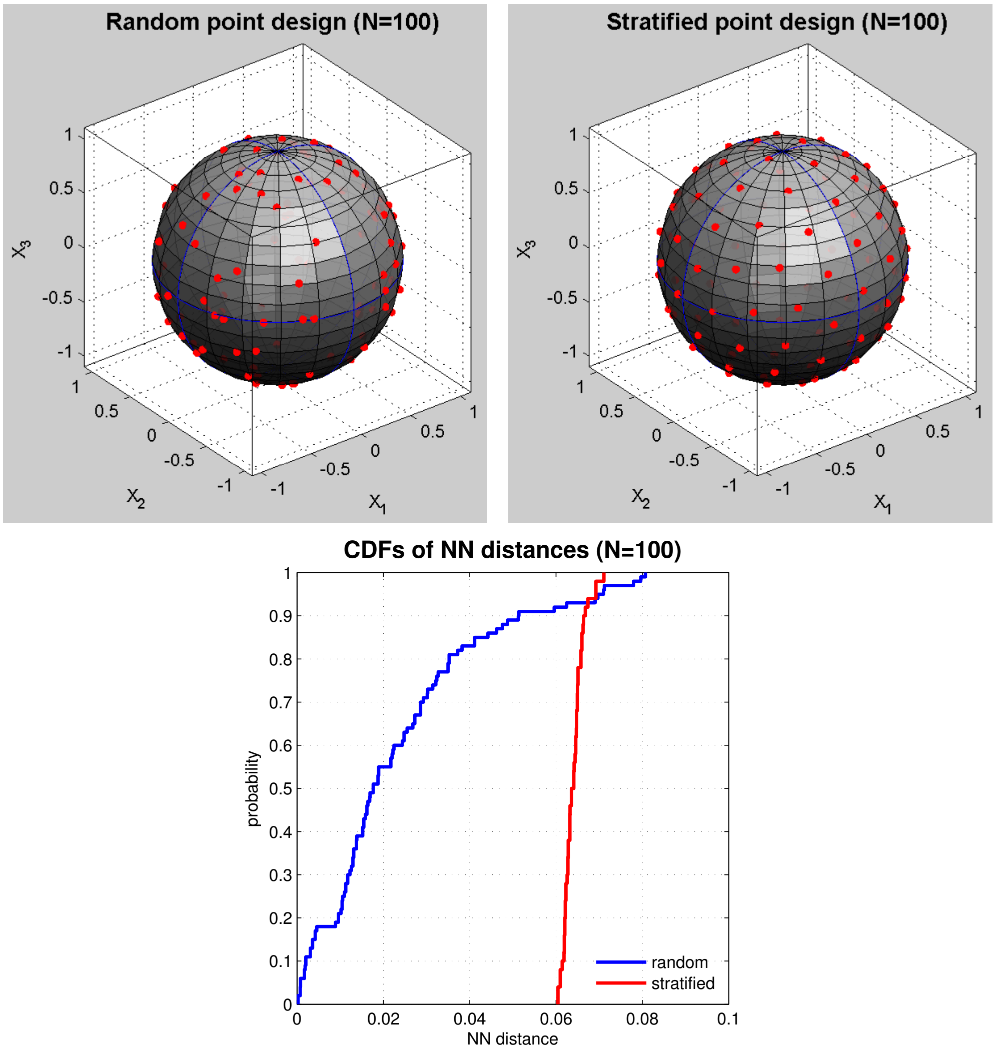


Figure 3.5: Random and stratified point design of 100 samples of 3 variables and the cdf of the nearest neighbor distances

1. Compute the (upper triangular) Cholesky factor $\mathbf{C}$ of the covariance matrix $\mathbf{\Sigma}_Y$ of the multivariate Gaussian PDF $\mathcal{G}(\boldsymbol{\mu}_Y, \mathbf{\Sigma}_Y)$ associated with $\mathcal{L}(\boldsymbol{\mu}_Z, \mathbf{\Sigma}_Z)$.

2. Generate a $(S \times M)$ matrix $\mathbf{V}_L$ with a stratified sample of size S from M uncorrelated standard Gaussian deviates; the entries of the columns of $\mathbf{V}_L$ can be generated using standard or midpoint LH sampling.

3. Compute the vector $\|\mathbf{V}_L\|$ with the S scalar norms of the rows of $\mathbf{V}_L$, and compute the $(S \times M)$ matrix $\tilde{\mathbf{U}}_L = diag(\|\mathbf{V}_L\|)^{-1}\mathbf{V}_L$; the S rows of $\tilde{\mathbf{U}}_L$ approximate S stratified, unit norm, realizations or points on the surface of a unit hyper-sphere in a S dimensional space.

4. Generate a vector $\mathbf{x}_L$ of S LH deviates from a chi distribution with M degrees of freedom, independent of $\mathbf{V}_L$, hence of $\tilde{\mathbf{U}}_L$; the S entries of $\mathbf{x}_L$, which can be also generated via standard or midpoint LH sampling, identify S representative hyper-ellipsoids of a M-variate Gaussian PDF via their corresponding Mahalanobis distances.

5. Scale/rotate and translate the entries of $\tilde{\mathbf{U}}_L$ as $diag(\mathbf{x}_L)\tilde{\mathbf{U}}_L\mathbf{C} + \mathbf{M}_Y$ so that they now define S, spatially correlated realizations with mean $\boldsymbol{\mu}_Y$, or S points lying on the surfaces of S different hyper-ellipsoids centered at $\boldsymbol{\mu}_Y$ and whose geometry is dictated by the eigenvalues and eigenvectors of the covariance matrix $\mathbf{\Sigma}_Y$; each such hyper-ellipsoid is indexed by the corresponding chi deviate in $\mathbf{x}_L$. This step is summarized as

$$\mathbf{Y}_L = \mathbf{W}_L\mathbf{C} + \mathbf{M}_Y = diag(\mathbf{x}_L)\tilde{\mathbf{U}}_L\mathbf{C} + \mathbf{M}_Y \tag{3.4}$$

   where $\mathbf{W}_L$ is a $(S \times M)$ matrix with stratified uncorrelated Gaussian deviates; Eq. (3.4) is a stratified version of Eq. (3.3).

6. Last, matrix $\mathbf{Y}_L$ can be transformed into a LH sample $\mathbf{Z}_L$ from a multivariate lognormal distribution as $\mathbf{Z}_L = \exp(\mathbf{Y}_L)$. The result is a set of S realizations from a lognormal random field, selected in a representative way from the associated likelihood model, thus ensuring that realizations are not too similar by chance as in simple random sampling.

A pictorial representation of generating S correlated samples on 3 points of a spatial setting via stratified likelihood sampling is depicted in Fig. 3.6.

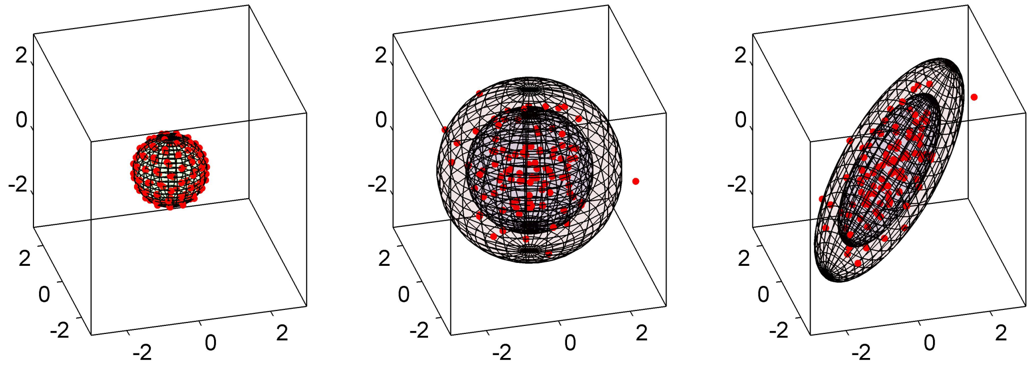


Figure 3.6: Stratified point design of S samples of 3 variables on a unit sphere $\tilde{\mathbf{U}}_L$, displacement on the new hyper-spheres of $diag(\mathbf{x}_L)$ radii, transformation/rotation to ellipsoids of different axes lengths according to $\mathbf{C}$

3.2 Accounting for model specific information

As stated in the introduction, the objective of efficient sampling is to construct a representative set of model responses using fewer – with respect to simple random (SR) sampling – realizations of model inputs, and consequently reduce the computational burden associated with model output uncertainty analysis due to multiple model evaluations. A major obstacle towards that objective for the stratified or controlled sampling methods presented above (likelihood sampling and Latin hypercube sampling) is that they operate exclusively on the space of model inputs. In other words, these methods capitalize on the statistics of model inputs only, and do not account for model-specific information, such as model sensitivity. Consequently, stratified likelihood or Latin hypercube sampling of model inputs need not necessarily lead to representative realizations of model outputs.

To mitigate this drawback, and along the lines of Switzer (2000), it is proposed that controlled sampling is conducted in a lower dimensional space – a smaller (than M) set of control locations – over which representative realizations of model inputs are generated. The specification of such locations is necessarily context-specific, and should be guided by the particular application at hand. When simulating radial flow and transport around an injection well and conductivity is assumed isotropic, for example, control points could be located radially away from the injection well, to account for the fact that concentration patterns are radially symmetric in this case. As suggested by Switzer (2000), one could also consider control points at regions of highest uncertainty in terms of data control, or alternatively in terms of model response uncertainty. The spatial pattern of such regions could be furnished from a preliminary analytical solution of the particular (e.g., initial and boundary value) problem at hand, when possible, or alternatively from a small set of model evaluations using attribute realizations generated via SR sampling.

An additional reason for selecting a relatively small number of control locations for the case of stratified likelihood sampling pertains to the exploration of the multivariate Gaussian distribution in terms of Mahalanobis distances and their associated chi distribution – the vector $\mathbf{x}$ of chi deviates in Eq. (3.3). More precisely, for a large number of simulation grid nodes M, the chi distribution with M degrees of freedom approaches a Gaussian distribution with mean $\sqrt{M-0.5}$ and variance equal to 0.5; see, for example Switzer (2000). This implies that the relative difference between a low and a high quantile of the chi distribution becomes smaller as M increases; which in turn implies that the exploration of the multivariate Gaussian likelihood of Eq. (3.3) becomes more concentrated and thus less informative; that is, one is simulating more realizations with a Mahalanobis distance close to $\sqrt{M-0.5}$; see also Bourgault (2012). Considering a lower dimensional multivariate Gaussian distribution via a small (less than 30) set of control locations, thus renders the efficient exploration of the associated likelihood function less concentrated (Switzer, 2000).

In what follows, a two-step, controlled sampling procedure is adopted to account for model sensitivity in stratified (log)Gaussian likelihood (and classical Latin hypercube) sampling. At the first step, attribute values at a small number of control locations are simulated using the stratified sampling methods described above. Stratified simulations at these few locations cover efficiently the range of possible attribute values and create a wider range of controlled variability at each control locations. At the second step, attribute values at the remaining locations are generated conditional on the fist step's simulated values. Representative realizations are thus generated across the entire simulation domain, due to the effects of spatial autocorrelation in attribute values.

More specifically, we denote as $\mathbf{y}_1$ the set of N simulated (and stratified) attribute values at a set of N control points $\{\mathbf{c}_n, n=1,\ldots,N\}$, and as $\mathbf{y}_2$ the set of simulated values at the remaining $M-N$ locations $\{\mathbf{c}_m, m=N+1,\ldots,M\}$. Given a set of N stratified simulated values $\mathbf{y}_1$, a conditional realization $\mathbf{y}_2$ can be generated as (Chilès and Delfiner, 1999)

$$\mathbf{y}_2 = \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{y}_1 + \mathbf{C}_{22|1}^{low}\mathbf{w}_2 \tag{3.5}$$

where $\mathbf{\Sigma}_{11}^{-1}$ is the $(N\times N)$ inverse covariance matrix between the N control points, $\mathbf{\Sigma}_{21}$ is the $((M-N)\times N)$ covariance matrix between the $M-N$ remaining simulation locations and the N control points; term $\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}$ defines the matrix of Simple Kriging weights given at the N control points for prediction at the $M-N$ remaining locations. Term $\mathbf{w}_2$ is a $((M-N)\times 1)$ vector of standard Gaussian deviates, and term $\mathbf{C}_{22|1}^{low}$ is the lower triangular Cholesky factor of the $((M-N)\times(M-N))$ conditional covariance matrix $\mathbf{\Sigma}_{22|1}$ defined as (Chilès and Delfiner, 1999)

$$\mathbf{\Sigma}_{22|1} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} \tag{3.6}$$

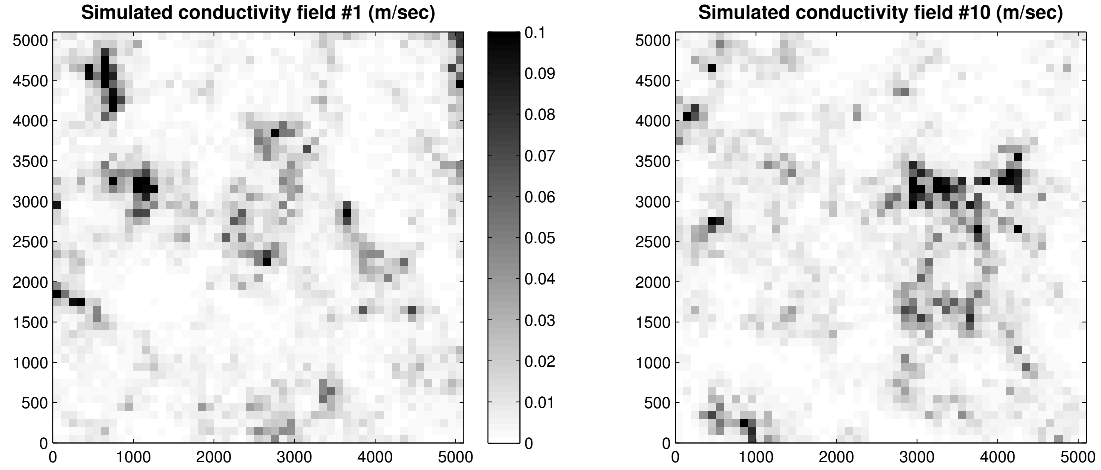
where $\mathbf{\Sigma}_{22}$ is the $((M-N)\times(M-N))$ covariance matrix between the $M-N$ remaining simulation locations, and $\mathbf{\Sigma}_{12}$ is the transpose of matrix $\mathbf{\Sigma}_{21}$ defined above. Term $\mathbf{C}_{22|1}^{low}\mathbf{w}_2$ in Eq. (3.5) represents a simulated realization – a SR sample of size 1 – of the Simple Kriging error via the Cholesky decomposition method of the conditional covariance matrix $\mathbf{\Sigma}_{22|1}$ (Chilès and Delfiner, 1999).

The above procedure can be repeated S times, resulting in a set of S stratified realizations over the entire simulation domain, per the stratification of the control values and the effects of spatial autocorrelation. Last, simulated realizations can be transformed to a lognormal distribution using the antilog transform to arrive at simulated realizations of hydraulic conductivity over the study region.

3.2.1 Two step - simulation case study

The efficiency of the various sampling methods described above in generating representative realizations from a (log)Gaussian random field model is evaluated within a hydrogeological context involving flow and transport in a porous medium, where the unknown spatial distribution of saturated hydraulic conductivity is modeled a as realization of a lognormal random field. More precisely, a simplified two-dimensional synthetic groundwater flow system is considered, similar to that used in Zhang and Pinder (2003). The dimensions of the flow system are 5100 m by 5100 m discretized into a 51 × 51 grid ($M = 2601$) with uniform rectangular cells of size 100 m by 100 m. Porosity is assumed constant throughout the domain and equal to 0.25. Flow boundary conditions consist of constant head of 0 m at the four corner cells and a constant head of 250 m at the central cell of the domain. No flow conditions ($\partial h/\partial n = 0$) are assigned to the rest of the domain boundaries, whereas the Modflow code (McDonald and Harbaugh, 1988) is used in this study to obtain the steady state flow solution.

A second-order stationary and isotropic lognormal random field is adopted with parameters borrowed from Sudicky et al. (2010). More precisely, the mean and variance of log conductivity are taken as $\mu_Y = -5.64$ and $\sigma^2_Y = 1.79$, respectively, corresponding to conductivity statistics $\mu_Z = 0.0087$ m/sec and $\sigma^2_Z = 0.00038^2$ (m/sec)^2. The semivariogram of log conductivity is assumed to be of exponential form, with no nugget effect, and effective range 1000 m, corresponding to one fifth of the domain extent along the cardinal directions. Two realizations of this random field model at the nodes of the domain described above are given in Fig. 3.7.


Simulated conductivity field #1 (m/sec)

Simulated conductivity field #10 (m/sec)

Figure 3.7: Two realizations of a lognormal random field modeling the spatial distribution of saturated hydraulic conductivity; see text for details.

For the solute transport problem, an initial concentration equal to 0 is assumed throughout the model domain. At time $t = 0$, a contaminant is introduced at the central cell, along the upstream constant head boundary, with constant concentration $C_0 = 100$ mg/l. No transport conditions ($\partial C/\partial n = 0$) are assigned along the domain boundaries. Longitudinal and transverse horizontal dispersivities are assumed to be equal to 5 m and

0.5 m, respectively. The MT3D code (Zheng, 1990) was then used to obtain breakthrough curves at the four observation wells located at the four corner nodes of the domain, as well as the solute transport solution up to an end time of $t = 2 \cdot 10^6$ sec; this end time represents a conservative time to a well-field failure and was selected, based on an initial set of simulations, as the maximum solute travel time to at least one observation well.

Reference ensemble statistics for solute concentration are derived from a set of 10000 solutions of the transport problem based on 10000 hydraulic conductivity realizations generated via simple random (SR) sampling. Two such concentration realizations derived from the conductivity realizations shown in Fig. 3.7 are given in Fig. 3.8.

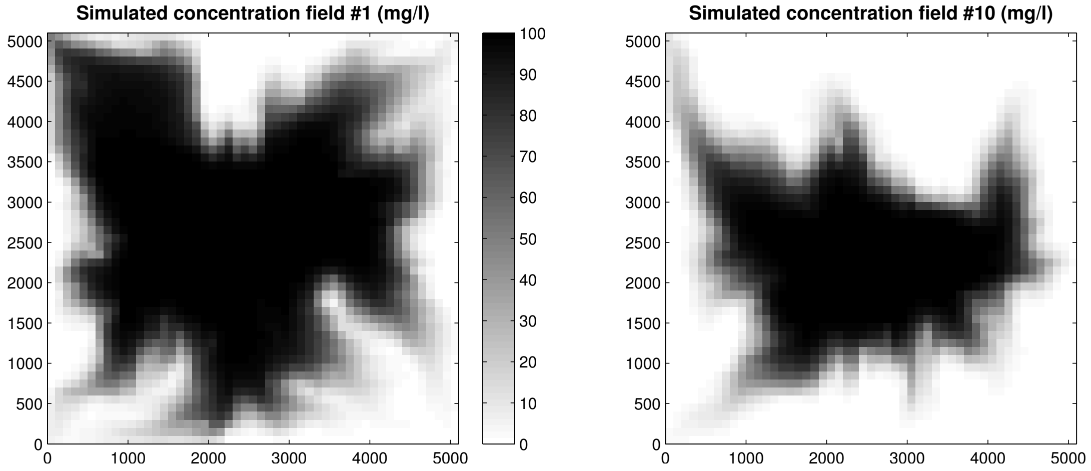


Figure 3.8: Two concentration realizations corresponding to the saturated hydraulic conductivity realizations of Fig. 3.7.

Reference statistics for solute concentration used for comparison of the various sampling methods considered in this work consist of: (i) the ensemble average field shown in Fig. 3.9A, (ii) the ensemble standard deviation field shown in Fig. 3.9B, (iii) the ensemble pairwise correlation between grid nodes – this involves all the entries of the upper triangular part of the concentration correlation matrix, (iv) the ensemble distribution of Mahalanobis distances between simulated concentration realizations and the ensemble average concentration field of Fig. 3.9A, and (v) the ensemble cumulative distribution function (CDF) of arrival times at 25% and 75% of the total concentration at any of the four observation wells.

The various sampling methods are compared in this work using a sample size equal to $S = 30$. Once a sample is generated, the discrepancy between the statistics of the simulated ensemble and the reference concentration statistics listed above is quantified using the root mean squared error (RMSE) for the case of summary statistics (mean, variance, correlation) or the mean absolute error (MAE) for the case of distribution percentiles (distribution of Mahalanobis distances). The computation of such error statistics is repeated over a set of $I = 100$ batches of realizations, each batch containing a sample size $S = 30$, thus estimating the sampling distributions of RMSE and MAE values for

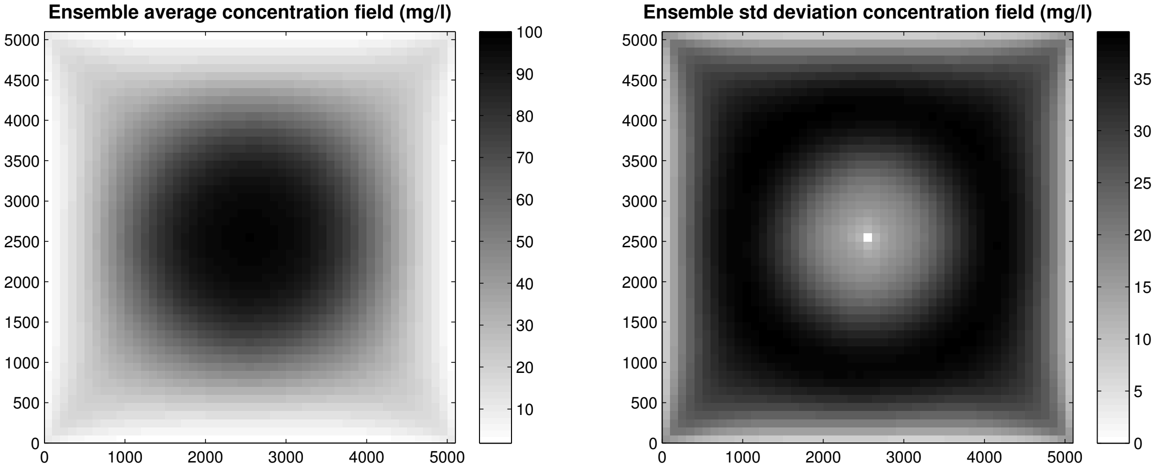


Figure 3.9: Reference ensemble average (A) and ensemble standard deviation (B) concentration fields computed from 10000 concentration realizations corresponding to 10000 hydraulic conductivity realizations generated via simple random sampling; see text for details.

each method; these distributions are presented in terms of their means and medians, as well as their 75% and 95% probability intervals. All subsequent figures, 75% RMSE (or MAE depending on the statistic selected) probability intervals are depicted with horizontal line segments, whereas 95% probability intervals with × symbols; median RMSE values are depicted as asterisks ($*$), whereas mean values as circles ($\circ$); horizontal dashed lines are also overlayed on the plots to facilitate visual comparison among the distributions. The better the reproduction of a reference statistic from simulations of a sampling method, the narrower the sampling distribution of the resulting, say, RMSE values, and the smaller (closer to 0) the center of that distribution.

3.2.2 Control point selection

In the proposed two-step sampling method accounting for model sensitivity, the generation of hydraulic conductivity fields consists of two steps. In the first step, efficient sampling methods (SL or LH) are employed to generate representative realizations of hydraulic conductivity at control points located within study region. In the second step, realizations of hydraulic conductivity are generated – using SR sampling – over the entire study region taking into account (conditional on) the representative samples generated in the first step. In what follows, an attempt is made to investigate the impact of control point selection (number of points and their location) on the quality of reproduction of the reference ensemble concentration statistics from the various sampling methods.

Control points should correspond to application-specific important locations, possibly controlling the variance of model outputs; in this case, the ensemble standard deviation of solute concentration of Fig. 3.9B. Control points are thus located radially away from the injection well, to account for the fact that concentration patterns were expected to be radially symmetric due to the isotropic random field. In addition, control points include locations in regions of large concentration standard deviation (doughnut-shaped area in Fig. 3.9B), and are varied in numbers from 1 to 25 points; see Fig. 3.10.

The central point of the study region is also included as an application-specific important point, since it corresponds to the location of the central injection well where the contaminant is first introduced in the region. Three more sampling schemes are considered for comparison, the two comprising a fine and coarse 4 × 4 grid (Fig. 3.11), and the third being a set of 16 random points changing locations in each of the 100 different iterations considered (Fig. 3.12).

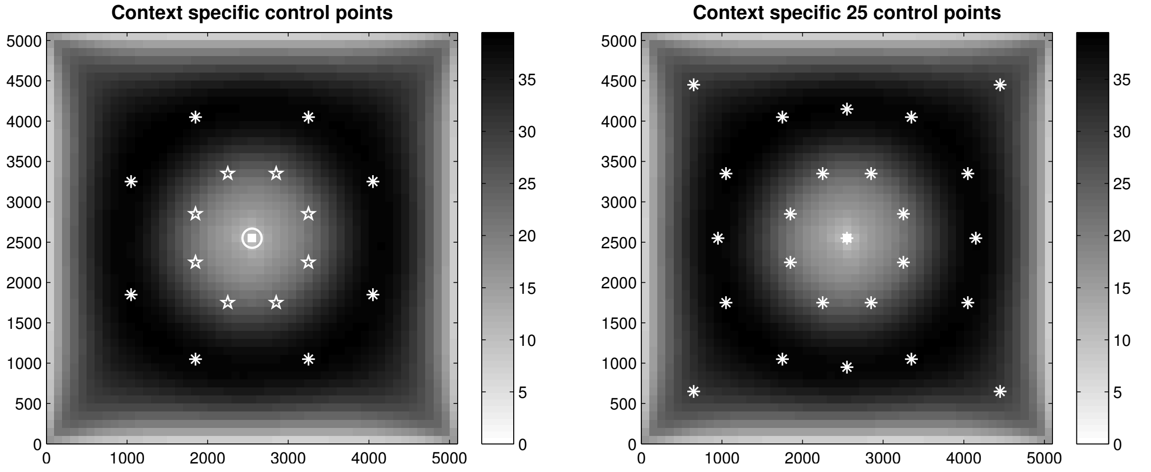


Figure 3.10: Two sets (16 points on the left, and 25 points on the right) of application-specific control points plotted on the ensemble standard deviation field of concentration.

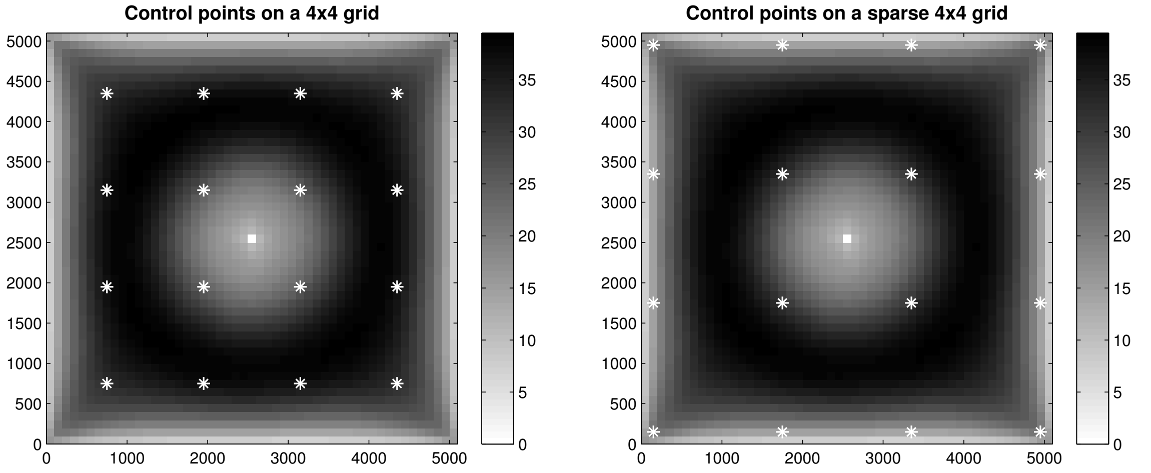


Figure 3.11: Two gridded sets of control points plotted on the ensemble standard deviation field of concentration: A) Control points chosen on 4 × 4 regular grid, B) Control points on a sparser 4 × 4 regular grid.

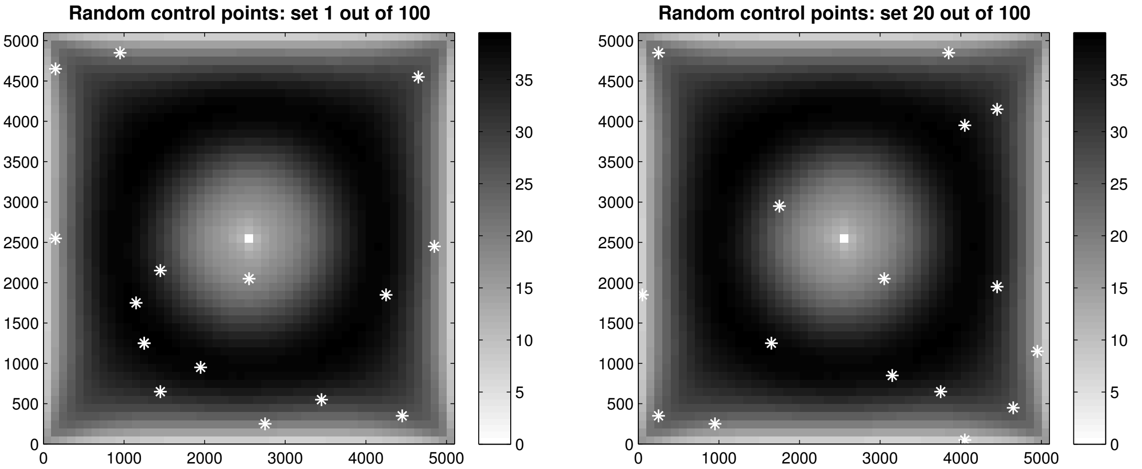


Figure 3.12: Two random sets of control points plotted on the standard deviation field of concentration; see text for details.

Overall, twelve different sets of control points are considered as candidates for fist-stage sampling, as follows:

- Only the central point of the region (circular marker in Fig. 3.10A), where the contaminant is first introduced to each model solution, denoted in the figures that follow as `src`.

- The central point along with 8 points radially emanating at the first areola (asterisk markers in Fig. 3.10A) corresponding to locations with concentration standard deviation values larger than $20mg/l$, denoted as `a1LHs` and `a1SLs` in the following figures; the first corresponds to LH and the second to SL sampling; `a1` denotes the first areola and `s` the source well.

- As above, but excluding the central point; denoted in the figures as `a1LH` and `a1SL`.

- Eight points radially emanating along the areola holding the largest values of standard deviation of the solute concentration field (star marker in Fig. 3.10A); denoted in the figures as `a2LH` and `a2SL`

- As previously, but including the central point; denoted in the figures as `a2LHs` and `a2SLs`

- The central point along with the points on the two areolae, excluding one point (different one for each iteration) from the second areola; denoted in the figures as `a12LH-` and `a12SL-`.

- Both of the above areolae excluding the central point(16 points in total); denoted in the figures as `a12LH` and `a12SL`.

- Sixteen points on a 4 × 4 regular grid (Fig. 3.11A); denoted in the figures as `g1LH` and `g1SL`.

- Sixteen points on a sparser 4 × 4 regular grid (Fig. 3.11B); denoted in the figures as `g2LH` and `g2SL`.
- Sixteen random points scheme: a different scheme is generated for each of the 100 iterations; denoted in the figures as `rndLH` and `rndSL`. Two random point schemes are displayed in Fig. 3.12.
- Twenty five control points including points in the two areolae, the central point, and four corner points near the observation wells (Fig. 3.10B); denoted in the figures as `25LH` and `25SL`.

All the above combinations of sampling schemes for the two sampling methods are compared in terms of their ability to reproduce the reference ensemble statistics computed via 10000 realizations generated via SR sampling. The results from above methods and control point locations are compared to the results obtained from 100 iterations of $S = 30$ SR samples. In all the figures that follow, the results of SR sampling are displayed at the end of each plot and denoted as `SR`.

3.2.3 Results

Figure 3.13 shows the reproduction of the ensemble average hydraulic conductivity field for the 22 combinations of control point schemes and the two stratified sampling methods (SL and LH), along with the results for simple random (SR) sampling. Reproduction is quantified here in terms of the sampling distribution of RMSE between reference and simulated ensemble conductivity statistics over the 51 × 51 grid cells. Since the reproduction ability for the different combination of sampling schemes and methods here is not conditioned to the application, the best reproduction depends only on the number of points and their dispersion over the study region. According to Fig. 3.13, the 25 sampling scheme, denoted as `25LH` and `25SL`, achieves the best reproduction of the reference mean conductivity field. Since there is no model-specific information, a better reproduction would be achieved by an equal number (25) of gridded points, where the stratification of control points would inform a larger area than the 25 points sampling scheme accounting for model sensitivity.

Figure 3.14 shows the reproduction of the ensemble average concentration field (Fig. 3.9A) for the 22 combinations of control point schemes and the two stratified sampling methods, along with the results for SR sampling. Reproduction is quantified here in terms of the sampling distribution of RMSE between reference and simulated ensemble concentration statistics over the 51 × 51 grid cells. One could easily state that the four control sampling schemes comprising the central point, apart from the one with only the second areola and the central point, surprisingly even the one with only the central control point, denoted as `src`, have a better reproductive ability than the other schemes. The scheme comprising the points at the second areola and the central point, denoted as `a2LHs` and `a2SLs` is not performing as good, due to the large distance between the central point and the second areola. Since there are no conditioning data in the area in between, i.e., the area through which the pollutant is transported at the beginning of each simulation, the central point does not play such a critical role in the simulation, thus not efficiently

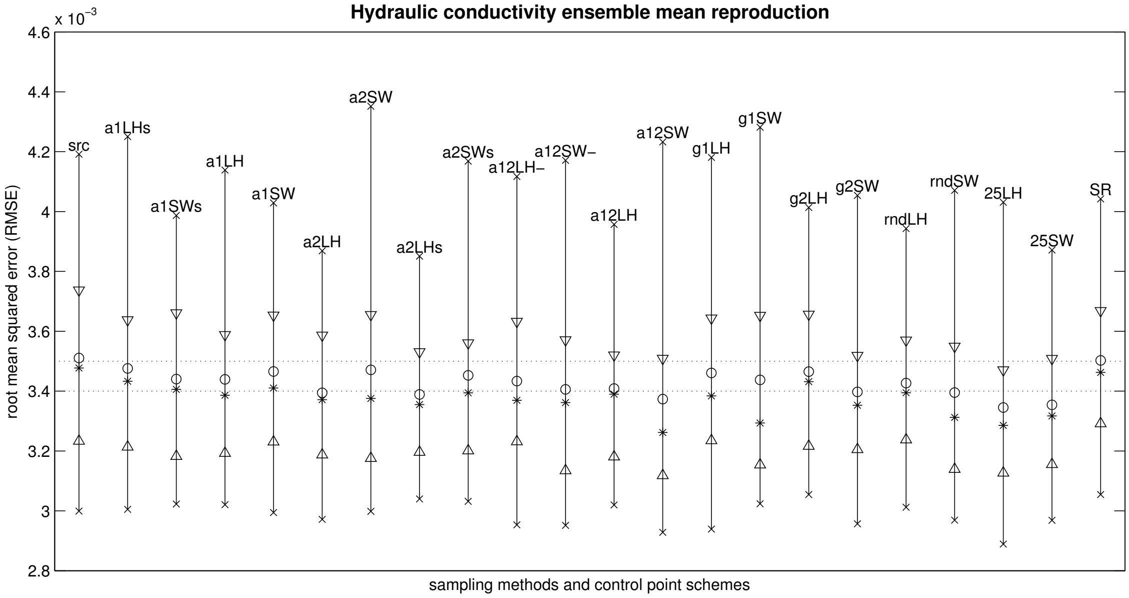


Figure 3.13: Reproduction of ensemble mean hydraulic conductivity field for various sampling schemes of control points; see text for details.

reproducing the reference ensemble mean. Moreover, all controlled sampling schemes yield better performance than SR sampling, and furthermore, no conclusive inference can be made with respect to the reproduction properties of the two sampling methods, LH and SL, across all sampling schemes.

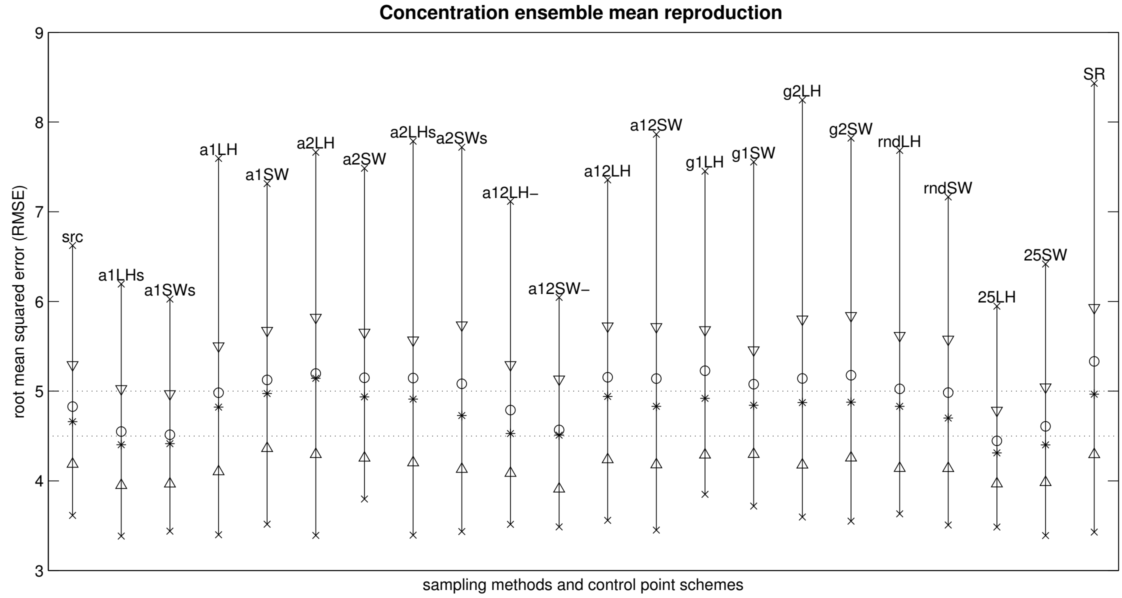


Figure 3.14: Reproduction of ensemble mean concentration field (shown in Fig. 3.9A) for various sampling schemes of control points; see text for details.

Figures 3.15 and 3.16 show the reproduction of the ensemble standard deviation concentration field (Fig. 3.9B) and the reproduction of the reference concentration correlation matrix (upper triangular part) for the 22 combinations of sampling methods and control point schemes. One could notice that here again the best reproduction is obtained via the sampling scheme with 25 application- specific control points, followed by LH on both areolae and the center point, and LH on first areola and center. All other methods or schemes yield approximately the same results, with few of them leading to worst reproduction than SR sampling.

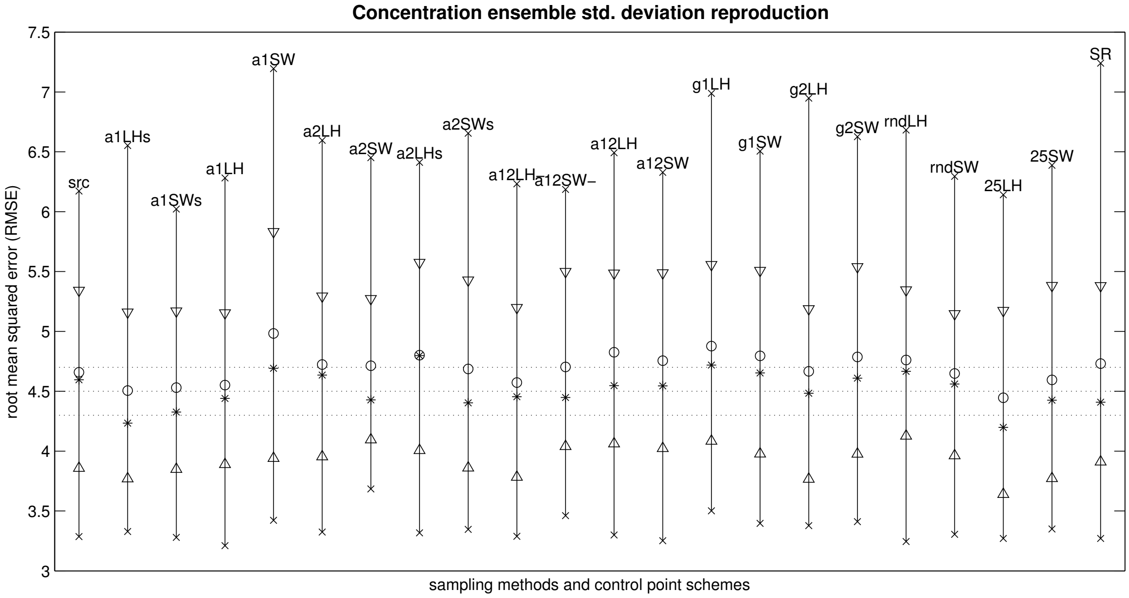


Figure 3.15: Reproduction of ensemble standard deviation concentration field (shown in Fig. 3.9B) for various sampling schemes of control points; see text for details.

Figure 3.17 shows the reproduction of the ensemble distribution of Mahalanobis distances between the simulated concentration realizations and the ensemble average concentration field (Fig. 3.9A). Reproduction is here quantified in terms of the sampling distribution of MAE computed between percentiles of the reference and simulated ensemble Mahalanobis distance distributions. All sampling schemes have approximately the same results, although one could notice the slightly better reproduction achieved by SL sampling compared to LH sampling, for the majority of control point schemes.

The effect of control point selection on the reproduction of the statistics of the solute concentration field over multiple time steps is investigated through Fig. 3.18 and Fig. 3.19. These figures show the reproduction of the reference cumulative distribution function (CDF) of arrival times for the 25% and 75%, respectively, of the total concentration at any of the four observation wells, and the corresponding CDFs obtained from the various combinations of sampling methods and schemes considered. Discrepancy between any two CDFs is quantified in terms of MAE between their corresponding percentiles, and the sampling distribution of this statistic is established over $I = 100$ batches of real-

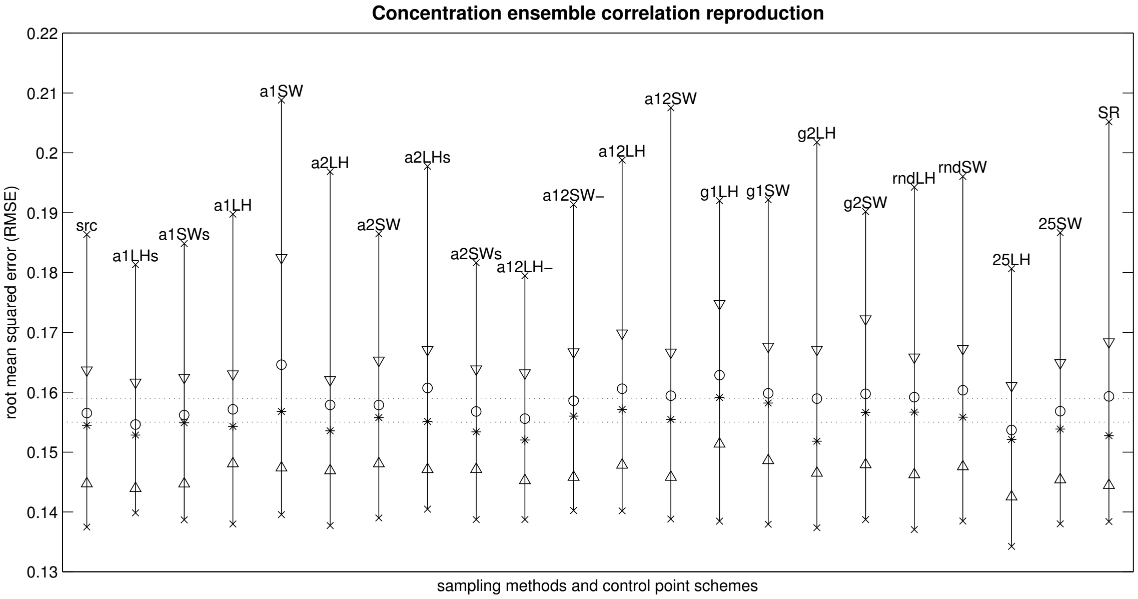


Figure 3.16: Reproduction of ensemble correlation matrix for various sampling schemes of control points; see text for details.

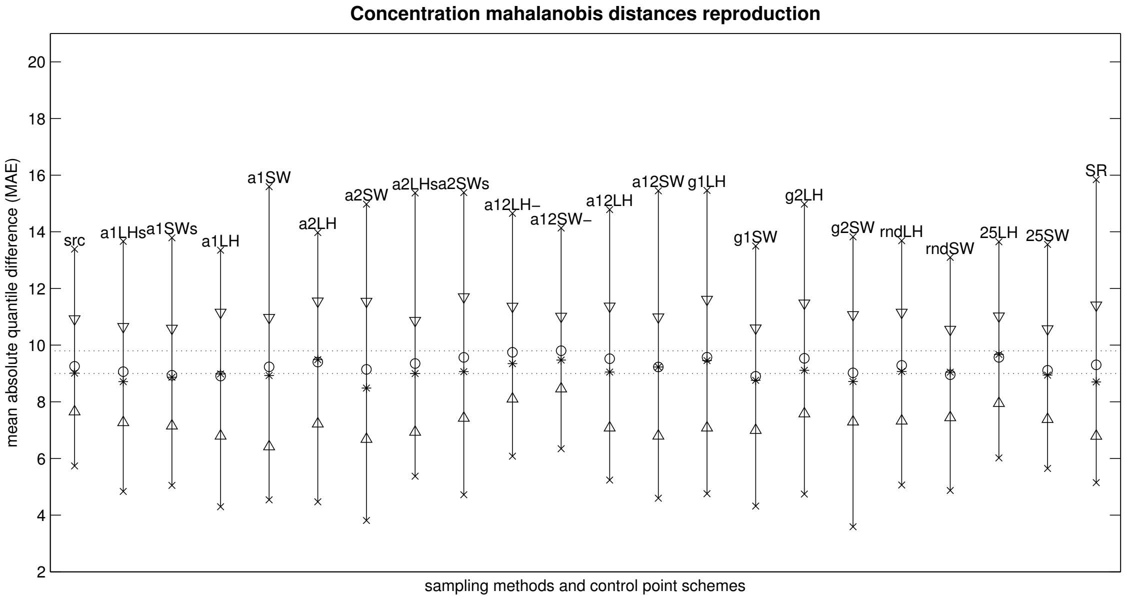


Figure 3.17: Reproduction of distribution of Mahalanobis distances between simulated realizations and the ensemble mean field, derived for various schemes of control points; see text for details.

izations as before. These last two figures investigate the efficiency of the various sampling methods and control point schemes considered at intermediate time steps corresponding to two proportions of total concentration for the observation well with the largest such concentration value. The three sampling schemes that include the central well and the first areola control points yield the best reproduction results.

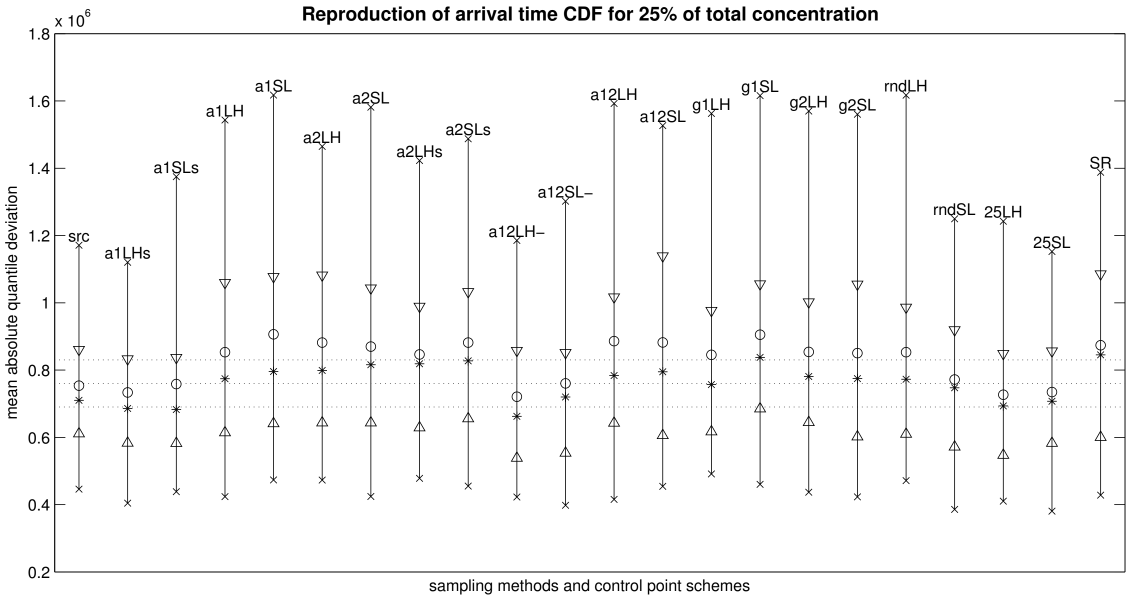


Figure 3.18: Reproduction of arrival times CDF for the 25% of total concentration, derived for various schemes of control points; see text for details.

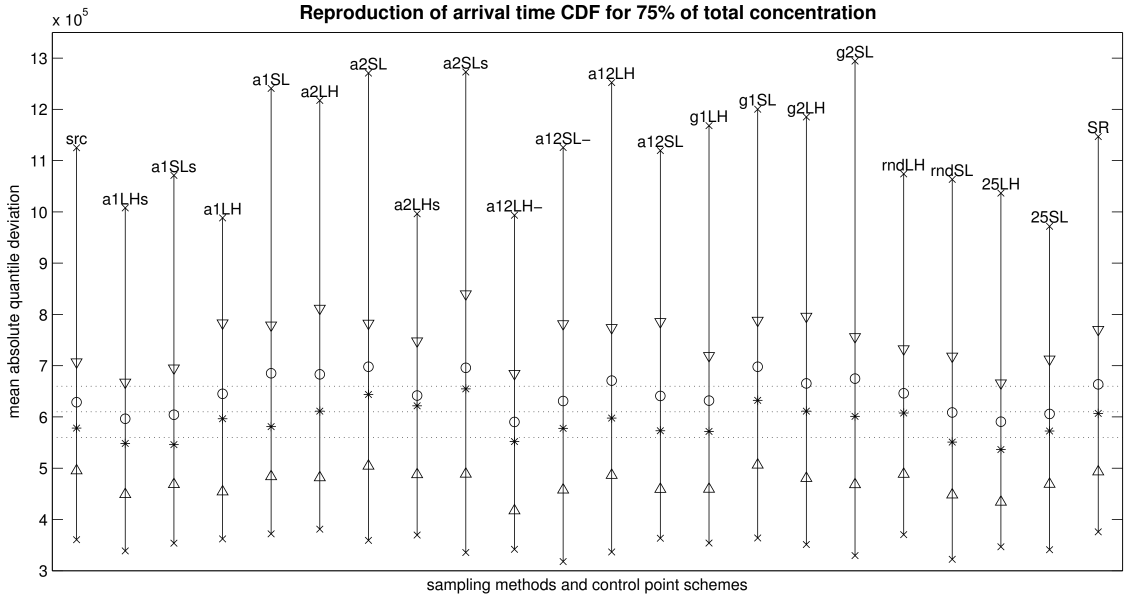


Figure 3.19: Reproduction of arrival times CDF for the 75% of total concentration, derived for various schemes of control points; see text for details.

3.3 Minimum energy sampling

As stated in Sec. 3.1 stratified likelihood sampling does not guarantee an optimal (with minimum energy) placing of the N points on the surface of a unit (hyper)sphere $\tilde{\mathbf{U}}_L$. In this section an alternative method of generating stratified points on the surface of the (hyper)sphere is proposed, named hereafter "Minimum Energy Sampling" (ME). ME sampling first generates N random points on the surface of the unit (hyper)sphere. Then a repulsive force vector, based on $1/r^2$, where r denotes the smallest linear distance between two neighbouring points, is calculated for each point. The resultant force vector is normalized, and then each point is displaced a distance $d = 1$ in the direction of that force, and finally projected back down onto the unit sphere (Bowman, n.d.). When the system nears convergence, the displacement vector for a given point is nearly in the same direction as the radius vector for that point due to the points being equally distributed (Hardin and Saff, 2004). The steps for generating N stratified points using ME sampling are shown in Fig. 3.20, where the point, depicts the transition from random points (small) to the system minimum energy convergence (large).

The ability of the three stratified sampling methods (LH-SL-ME) considered in this study, at furnishing representative attribute values, in terms of maximum dissimilarity between them, was explored by generating 100 batches of correlated hydraulic conductivity values at nine control points of a case study domain similar to that of Sec. 2.4.1. This case study only focuses in the exploration of the dissimilarities of the simulated values at the control points and how they relate to the ability of the sampling method to reproduce values covering the largest possible range of the attribute distribution; the larger discrepancies of values indicate highest coverage of the distribution range by the sam-

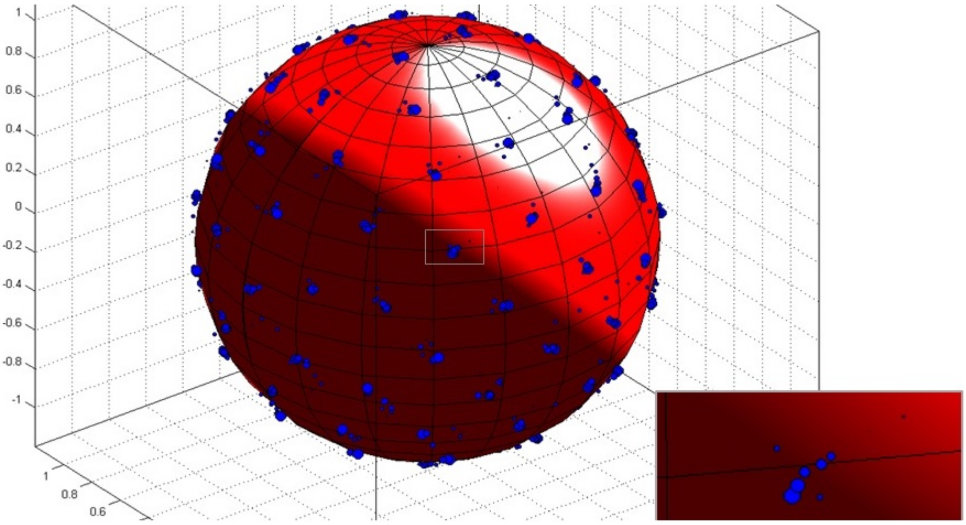


Figure 3.20: Three dimensional depiction for generating a sample of N points on the surface of the unit (hyper)sphere

pling method, thus a better reproductive ability with fewer realizations. The results are illustrated in Fig. 3.21 (top), where for both sample sizes under consideration −10/30−, ME sampling displays the largest nearest neighbor dissimilarities distribution (expressed in terms of the relevant cumulative distribution function) of the simulated hydraulic conductivity values. Figure 3.21 (bottom) depicts the dissimilarities of concentration values at the same control points, resulting from the hydrogeological model evaluation similar to that of Sec. 2.4.2. In this case, a slightly better reproductive ability (larger nearest neighbor distances) can be distinguished for ME sampling especially in the left graph referring to the 10 realizations sample size.

3.3.1 Two step simulation example

The above methods, along with the LH sampling method are applied in a dimensionality reduction context by selecting flow-controlling points over which representative sampling of hydraulic conductivity is performed, thus also accounting for the sensitivity of the flow and transport model to the input hydraulic conductivity field (Caers, 2011). According to Kyriakidis and Gaganis (2013), one could consider control points at regions of highest uncertainty in terms of data control, or alternatively in terms of model response uncertainty. Moreover, control points should correspond to application-specific important locations in terms of controlling the variance of realizations of model outputs; in this case, the ensemble standard deviation of solute concentration (Fig. 3.22).

Similarly to Sec. 2.4, the performance of sampling methods, LH, SL, and ME, is compared for different sample sizes $N(10-30)$, to that of SR sampling in terms of reproduction of ensemble statistics of reference fields; a set of 10000 SR realizations hydraulic conductivity and solute concentration fields. More specifically, the same 51 × 51 simula-

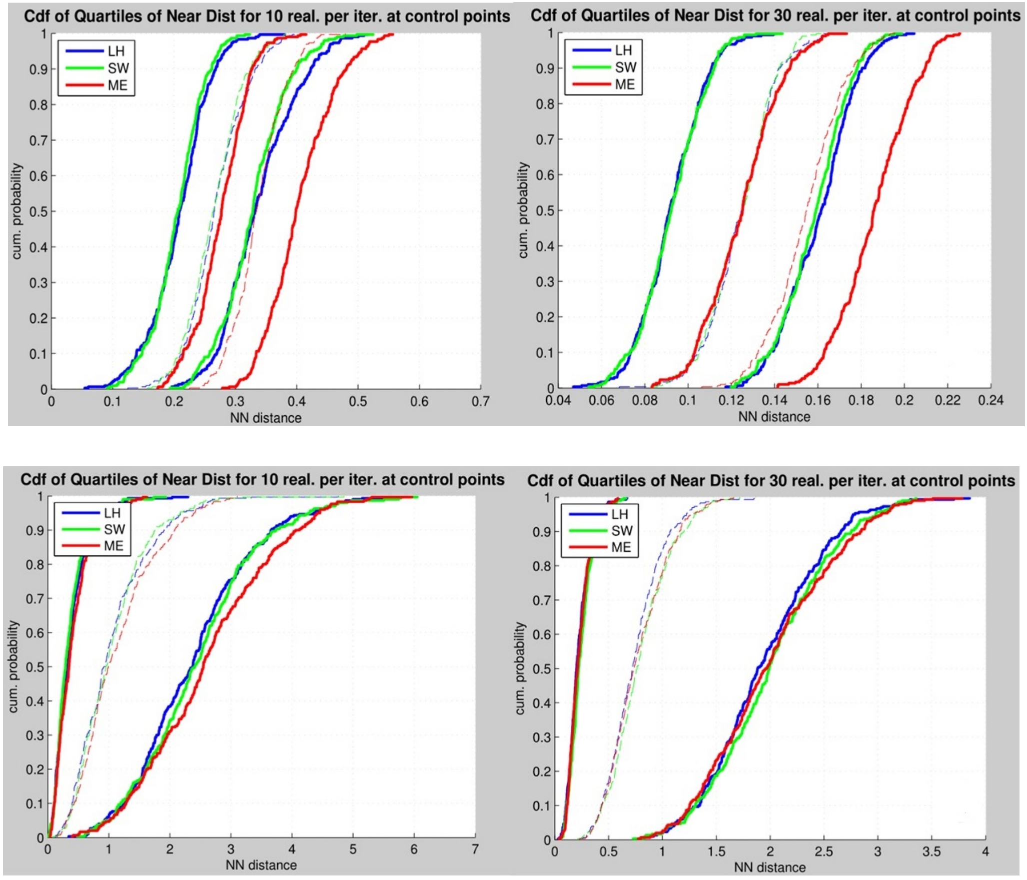


Figure 3.21: $5^{th}, 50^{th}$, and 95^{th} percentiles of CDF of nearest neighbor distances dissimilarities between simulated values of hydraulic conductivity (top) and concentration (bottom), at nine control points (eight points radially emanating along the first areola depicted as stars along with the central point depicted as circle in Fig. 3.22).The metrics considered are: (i) cosine, for hydraulic conductivity (top) and (ii) mahalanobis, for concentration (bottom).

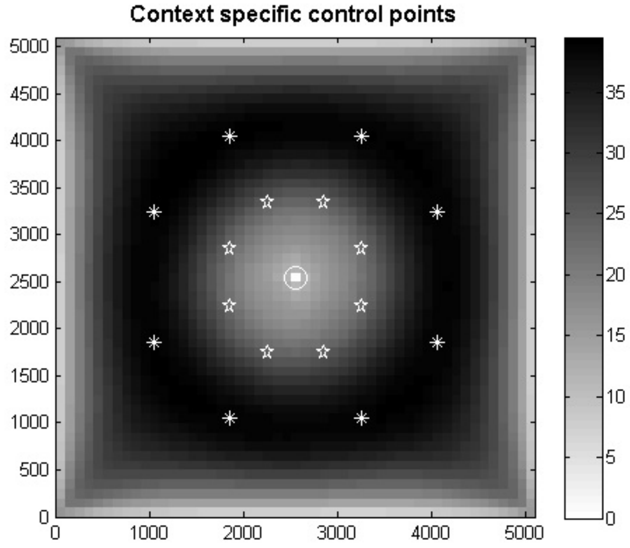


Figure 3.22: Application - specific control points plotted on the ensemble standard deviation field of concentration

tion grid was adopted. Once a sample, say of size $S = 10$, is generated, the discrepancy between the statistics of the simulated ensemble and the reference conductivity statistics listed above is quantified using the root mean squared error (RMSE) for summary statistics (mean, standard deviation, correlation and Mahalanobis distances).The computation of such error statistics is repeated over a set of $I = 100$ batches of realizations, with each batch containing the same sample size, $S = 10$ for example, thus estimating the sampling distributions of RMSE values for each sample size and for each method; these distributions are presented in terms of their means and medians, as well as their 75% and 95% probability intervals. The better the reproduction of a reference statistic from simulations of a sampling method with a given sample size, the narrower the sampling distribution of the resulting, say, RMSE values, and the smaller (closer to 0) the center of that distribution. Fig. 3.23 illustrates the ensemble mean and standard deviation fields of the 10000 reference SR realizations.

3.3.2 Results

The performance of the proposed ME sampling method was investigated in a hydrogeological context via a simple case study involving flow and transport in a heterogeneous porous medium, in comparison to stratified sampling methods LH and SL along with SR sampling. The statistics considered for hydraulic conductivity included the ensemble mean, standard deviation, as well as short-scale correlation. The reproduction of the ensemble mean, standard deviation and distribution of Mahalanobis distances from the ensemble mean for the ensemble concentration field resulting from solving a flow and transport boundary problem for each hydraulic conductivity realization, was also evaluated in the second part of the case study. For all statistics considered for both model

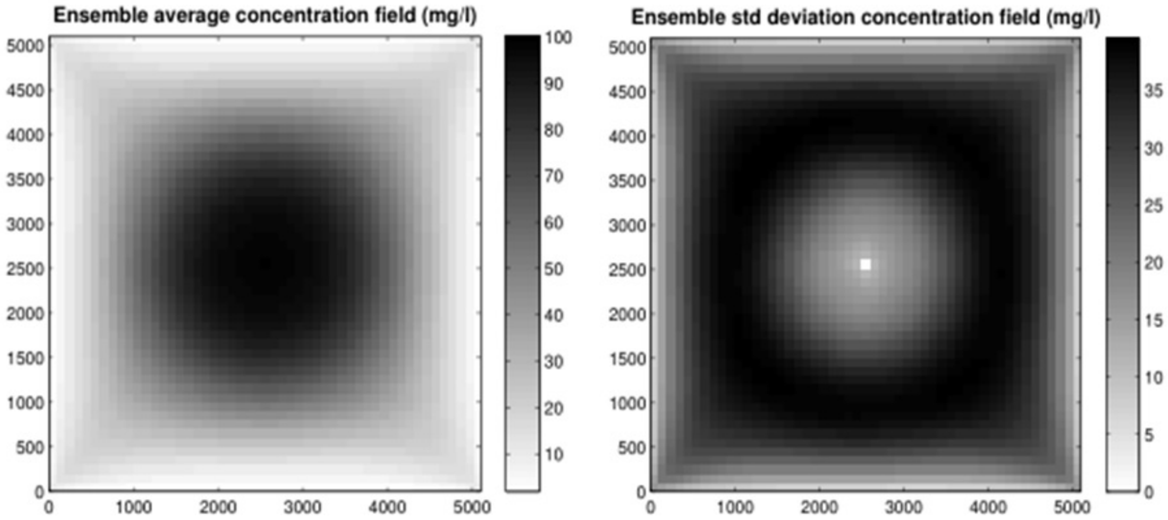


Figure 3.23: Ensemble mean (left) and standard deviation (right) reference concentration fields

inputs (Fig. 3.24) and outputs (Fig. 3.25), ME sampling constitutes an equal if not more efficient simulation method than LH and SL sampling, as it can reproduce to a similar extent statistics of the reference conductivity and concentration fields, yet with a slightly smaller sampling variability than SR sampling. Concluding, the proposed ME sampling method offers a viable alternative to existing stratified sampling methods.

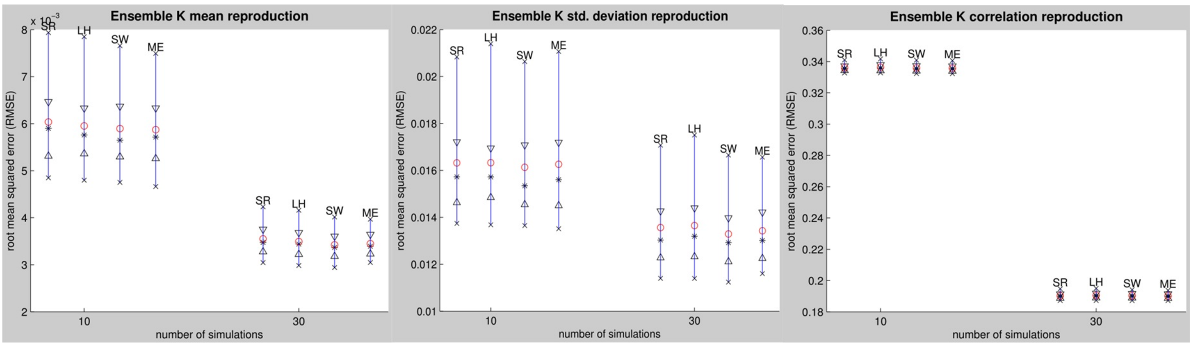


Figure 3.24: Reproduction of ensemble statistics i) mean, ii) std deviation and iii) correlation, between hydraulic conductivity realizations and the ensemble average hydraulic conductivity field. Reproduction is quantified here in terms of the sampling distribution of RMSE between reference and simulated ensemble concentration statistics

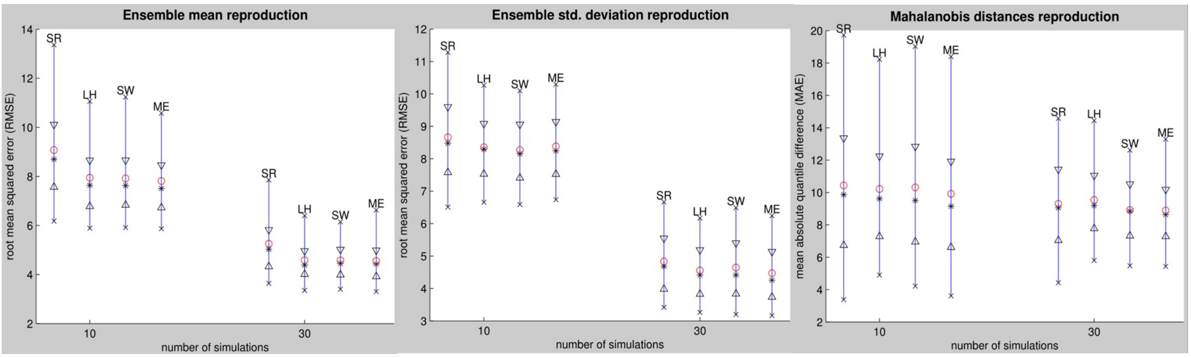


Figure 3.25: Reproduction of ensemble statistics i) mean, ii) std deviation and iii) mahalanobis distances, between concentration realizations and the ensemble average concentration field

3.4 Conclusions and Discussion

This Chapter adopted two different methods of stratified sampling, based on different properties than Latin Hypercube, and applied these methods for the first time in a spatial (and hydrogeological) context. Moreover, the concept of two step sampling was further employed for the simulation procedure, thus accounting for the uncertainty of the particular case study at hand. The novelty of the proposed approaches are summarized below along with potential future methodological and application related extensions.

Methods Uncertainty analysis is often employed in earth and environmental sciences applications for evaluating the impact of spatial uncertainty or variability in model parameters to model predictions. In many applications, the spatial distribution of model parameters is often conceptualized in terms of a (log)Gaussian random field. Realizations from such a random field are typically generated via simple random (SR) sampling, and are then used through repeated model evaluations in a Monte Carlo framework for assessing the uncertainty in the spatial distribution of model predictions. Deriving a representative set of model predictions, however, is often computationally demanding due the need for repeated complex model evaluations.

Three controlled or stratified simulation methods, namely stratified likelihood (SL), minimum energy (ME) and Latin hypercube (LH) sampling, were considered in this work for generating a smaller set of more representative – than SR sampling – attribute realizations (spanning or covering more efficiently the range of possible values) of model inputs; hence, it is expected, by extension, of model outputs. Controlled sampling of model inputs, however, capitalizes on the statistics of those inputs only, and does not account for model-specific information, such as model sensitivity. Consequently, controlled sampling of model inputs need not necessarily lead to representative realizations of model outputs. To address this lack of application or model-specific information, a two step simulation procedure was proposed in this work: in the first step, representative samples are generated at control sites over the study region; model input values at these control sites should be most influential on model predictions. These simulated attribute values are subsequently used as conditioning data for generating, via simple random sampling, attribute realizations over the remaining set of simulation locations.

Case Studies The performance of the proposed two-step simulation procedure for controlled sampling was investigated in a hydrogeological context via two synthetic case studies involving flow and transport over a simple geometrical domain, assuming known initial and boundary conditions. Simulated hydraulic conductivity fields were generated using multiple control point schemes, and three stratified sampling methods (SL, ME and LH) in addition to SR sampling. The simulated conductivity fields were utilized in a Monte Carlo framework to conduct simulations of solute transport and compute the associated concentration fields. The end time for the solute transport solution was set to $2 \cdot 10^6$ sec, representing a conservative time to a well-field failure; this end time was selected as the maximum solute travel time to at least one observation well, based on an initial set of model evaluations. Reference ensemble concentration statistics were computed by evaluating the flow and transport model using a very large number – 10000 – of conductivity realizations generated via SR sampling. The concentration statistics considered included the ensemble mean, standard deviation, as well as correlation and distribution of Mahalanobis distances from the ensemble mean field.

To investigate the sensitivity of concentration statistics reproduction on control point specification, many combinations of such points were considered in the case study of Sec. 3.2.1, varying both the number and their location, accounting for (Fig. 3.10) and ignoring (Fig. 3.11 and Fig. 3.12) model sensitivity. In the former case, control points were located radially away from an injection well and in regions of highest uncertainty in terms of expected model response. Flow and transport, however, is a dynamic process, and consequently the location of such regions of higher uncertainty is inevitably tied to the simulation end time ($2 \cdot 10^6$ sec); this implies that the specification of control points also depends on that end time. If one were to account for the time varying nature of the model in this particular application, control points should only be located radially away from the central well, not targeting any elusive region with (expected) higher uncertainty.

Overall, two-step controlled sampling performed better than simple random sampling; the best reproduction of ensemble statistics (except for the Mahalanobis distance for which all methods yield similar results) was achieved by the sampling scheme denoted as `25LH` and `25SL` consisting of 25 control points (Fig. 3.10), followed by the 16 and 9 control points schemes including the central well, denoted as `a12LH-`, `a12SL-` and `a1LHs`, `a1SLs`, respectively. Schemes that excluded the central (most influential) point corresponding to the injection well location did not effectively reproduce ensemble concentration statistics. Similarly, sampling schemes that included points in the first areola (closer to the central well) were in general more efficient. Moreover, point schemes that did not account for model-sensitivity information, i.e., the random and gridded point sets of Fig. 3.12 and Fig. 3.11, did not efficiently reproduce the reference statistics. The second most important feature of control sampling was the number of control points, since the order of the three better reproducing point schemes was found to be regulated by this factor, following an "efficiency order" of $25 - 16 - 9$ points. LH sampling involving all $M = 2601$ simulation locations, i.e., not taking into account any model sensitivity, resulted in a very close reproduction of ensemble statistics (mean and variance) of the concentration field for this particular application; the corresponding error statistics exhibited the smallest variability in their sampling distributions across all controlled methods considered. It is argued that such a close reproduction of ensemble statistics might be deemed undesirable, as it might provide rather narrow bounds of model output variability.

Conclusions and future work Controlled sampling from (log)Gaussian random fields (pertaining here to hydraulic conductivity) accounting for model (here of flow and transport) sensitivity was found in this work to be more efficient than simple random sampling, i.e., reproduced better the ensemble statistics of model predictions (here of solute concentration for a simplified transport problem). For highly non-Gaussian random field models and for highly non-linear transfer functions, the uncertainty modeling framework advocated by (Caers, 2011), which calls for a proxy model to evaluate the links between model inputs and outputs (including sensitivity to initial and boundary conditions), provides a more general framework to efficient uncertainty analysis.

For the (log)Gaussian case and the simplified flow and transport problem considered in this work, effective first step simulation control provided reasonable control of the entire simulation due to the effects of spatial autocorrelation. Evidently, the careful selection of key locations whose attribute values are expected to exert significant impact on model predictions, is critical for controlling sampling variability across multiple simulations. Such a selection is necessarily context-specific, and should be guided by the particular application at hand. Control points were here located radially away from an injection well, to account for the fact that concentration patterns were expected (due to the isotropic random field) to be radially symmetric. Control points also included locations in regions of highest uncertainty in terms of model response, tied to the particular design of the simulation experiment adopted in the case study. The spatial pattern of such regions could be furnished from a preliminary analytical solution of the particular problem at hand, when possible, or alternatively from a small set of model evaluations using simulated attribute realizations generated via simple random sampling.

In a real case hydrogeological flow and transport problem scenario, control points could be guided by the available independent parameters related to each particular problem. Potentially available seismic data could guide the user's selection of uncertain regions, based on the distribution range of data per location; that is, a larger range of historic seismic data indicating a higher uncertainty in the presumed hydraulic conductivity values. The complexity of geological features could also indicate regions of higher uncertainty; the more complex the geological formations, the higher uncertainty in the presumed hydraulic conductivity fields. Both of the above examples and any other multivariate analysis could lead the relevant stakeholders to locations where additional geological samples should be extracted from in order to minimize the underlying uncertainty. Since sampling is highly cost related and parameters related to hydrogeological problems are continuous in space additional sampling always meets the limitation of not examining the underground as exhaustively as a researcher might want. Control sampling could really help on that front by choosing locations of high uncertainty, potentially close to locations of additional samples. This way, the uncertainty analysis simulation procedure would ensure a more efficient quantification of uncertainty by combining information from additional samples with the stratified simulation on control points, both located on areas of high uncertainty. This combination is offering to the modeler, low computational cost by performing the fewest possible additional samples, and high performance by employing stratified simulation in well targeted locations of high uncertainty.

Chapter 4

Latin Hypercube Sampling on Large Grids

In this chapter, a novel Latin Hyperube sampling method is proposed for efficiently generating realizations of lognormal random fields on very large (order of millions of nodes) regular grids. LH sampling methods typically rely in practice on the Cholesky decomposition of a covariance matrix pertaining to all possible pairs of simulation grid nodes (Zhang and Pinder, 2003). Consequently, the routine application of LH sampling is computationally prohibitive for large simulation grids, since it becomes difficult to store it in computer memory and even more difficult (time consuming) to compute its Cholesky decomposition. In other words, for large grids, which unfortunately is the typical case in a hydrogeological context (Chilès and Delfiner, 1999), simulation via the Cholesky factor of the covariance matrix (Eq. 2.23) becomes prohibitive, due to storage requirements and computational costs associated with the covariance matrix decomposition (Chilès and Delfiner, 1999). Although there exist simulation methods for generating realizations of random fields on large regular grids, these methods have not been employed to date within a Latin hypercube sampling context.

In a nutshell, the proposed methodology combines an existing method of Latin hypercube sampling Stein (1987) with an efficient simulation method for simple random sampling of Gaussian (or transformed Gaussian) random fields with stationary covariance functions (Dietrich, 1997). The result is a practical and efficient methodology for extending LH sampling on large grid domains, thus promoting the use of uncertainty analysis techniques in practice for potentially real-life models with spatially distributed parameters.

The remainder of the chapter is structured as follows: Section 4.1 describes the proposed method of LH sampling of Gaussian or transformed to be Gaussian random fields that can handle simulation on very large grids with possibly millions of nodes, and Section 4.2 presents two high discretization (a $2D$ and one $3D$) synthetic case studies involving flow and transport in a heterogeneous porous medium illustrating the benefits of the proposed Latin hypercube over simple random sampling.

4.1 Latin hypercube simulation of random fields on large grids

In this Section, a novel method is developed for Latin hypercube (LH) sampling of second-order stationary Gaussian random fields on large regular grids. Subsection 4.1.1 provides the motivation for developing the proposed LH sampling method, Subsection 4.1.2 describes the computational properties of covariance matrices derived from second-order stationary covariance models evaluated at distances corresponding to pairs of nodes of a regular grid, whereas Subsection 4.1.3 presents the procedure for LH simulation of Gaussian fields on large regular grids and provides some pertinent examples.

4.1.1 Introduction

A solution to this aforementioned computational problem related to the Cholesky decomposition is proposed in this Chapter, based on the fact that LH sampling is essentially a post-processing method that transforms a spatially correlated simple random (SR) sample into a marginally stratified correlated LH sample; see Sec. 2.3.2. The key observation is that Stein's LH sampling method is independent of the simulation algorithm used to generate the SR sample. Hence, LH sampling is not tied to the Cholesky decomposition of a possibly large covariance matrix and can be in principle applied for LH simulation on very large grids, as long as a suitable method is available to generate SR samples at those grids.

In what follows, we develop a novel LH simulation method that is capable of generating LH samples from second-order stationary Gaussian random fields on very large regular grids (order of millions of nodes). In a nutshell, the proposed LH sampling methods amounts to combining an existing method for generating realizations of second-order stationary random fields on large regular grids using simple random (SR) sampling (Dietrich, 1997) with the LH sampling method of Stein (1987).

The SR sampling algorithm used in the proposed LH sampling method assumes a multivariate Gaussian PDF $\mathcal{G}(\mathbf{0}, \mathbf{\Sigma})$ and is based on the square root decomposition of the covariance matrix $\mathbf{\Sigma}$ given in Eq. (2.11). A $(M \times 1)$ correlated Gaussian random vector is then generated from a $(M \times 1)$ uncorrelated standard Gaussian random vector $\mathbf{w}$, as

$$\mathbf{y} = \sqrt{\mathbf{\Sigma}}\mathbf{w} \tag{4.1}$$

where $\sqrt{\mathbf{\Sigma}}$ the $(M \times M)$ square root of the covariance matrix $\mathbf{\Sigma}$; note that subscript Y is dropped from $\mathbf{\Sigma}_Y$ for simplicity.

Since $\sqrt{\mathbf{\Sigma}} = \mathbf{Q}\sqrt{\mathbf{E}}\mathbf{Q}^T$, Eq. (4.1) becomes

$$\mathbf{y} = \mathbf{Q}\sqrt{\mathbf{E}}\mathbf{Q}^T\mathbf{w} = \mathbf{Q}[diag(\sqrt{\mathbf{e}})]\mathbf{Q}^T\mathbf{w} = \mathbf{Q}\left(\sqrt{\mathbf{e}} \odot (\mathbf{Q}^T\mathbf{w})\right) \tag{4.2}$$

where $\odot$ denotes element-by-element multiplication or the Hadamard product. Equation (4.2) implies that a vector $\mathbf{y}$ of M correlated Gaussian deviates can be generated from a vector $\mathbf{w}$ of uncorrelated standard Gaussian deviates by: (i) computing the matrix-vector product $\mathbf{Q}^T\mathbf{w}$; the result is a $(M \times 1)$ vector, (ii) computing the Hadamard product of $\sqrt{\mathbf{e}}$ and $\mathbf{Q}^T\mathbf{w}$; the result is a $(M \times 1)$ vector, and (iii) finally computing the matrix-vector product $\mathbf{Q}\left(\sqrt{\mathbf{e}} \odot (\mathbf{Q}^T\mathbf{w})\right)$ to derive the resulting $(M \times 1)$ vector $\mathbf{y}$.

At a first glance, the simulation method for SR sampling of Eq. (4.2) seems computationally inefficient, since computing the eigenvectors and eigenvalues of an arbitrary covariance matrix $\mathbf{\Sigma}$ is a more expensive operation than the Cholesky decomposition. Fortunately, however, there is a very particular (albeit very useful) case whereby the computation of eigenvectors and eigenvalues of $\mathbf{\Sigma}$ can be done very efficiently. This is the case of a stationary covariance model evaluated for separation distances corresponding to distances between the nodes of a regular grid.

4.1.2 Computationally exploitable covariance

When considering M RVs corresponding to the M nodes of a regular grid and a stationary covariance model $\sigma(\mathbf{h})$, the structure of the covariance matrix $\mathbf{\Sigma}$ is highly exploitable from a storage and computational perspective (Zimmerman, 1989). More precisely, in 1D, $\mathbf{\Sigma}$ is a symmetric Toeplitz matrix and is completely characterized only by its first row (or column). In higher dimensions, $\mathbf{\Sigma}$ is a symmetric block Toeplitz matrix with Toeplitz blocks and is completely characterized only by its first block row.

For a regular 1D grid of M nodes with unit spacing, for example, the corresponding covariance matrix $\mathbf{\Sigma}$ becomes

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma(0) & \sigma(1) & \cdots & \sigma(M-2) & \sigma(M-1) \\ \sigma(1) & \sigma(0) & \cdots & \sigma(M-3) & \sigma(M-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma(M-2) & \sigma(M-3) & \cdots & \sigma(0) & \sigma(1) \\ \sigma(M-1) & \sigma(M-2) & \cdots & \sigma(1) & \sigma(0) \end{bmatrix}$$

and is completely characterized by its first row vector $[\sigma(0) \cdots \sigma(M-1)]$. The $(j+1)$-th row of the above array is constructed by shifting the preceding j-th row one position to the right, and filling the leading empty position in the $(j+1)$-th row with the j-th element of vector $[\sigma(1) \cdots \sigma(M-1)]$.

Even more computationally exploitable covariance matrices arise when considering a stationary covariance model $\sigma(\mathbf{h})$ evaluated at distances corresponding to the distances between the $\tilde{M} = 2M$ nodes of an embedded regular grid with *toroidal* boundaries; that is a "wrapped" covariance matrix whose lower right value (block of values in 3D) is "neighbor" to the upper left value. Smaller embeddings (with $\tilde{M} < 2M$) are possible, depending on the range of the covariance model $\sigma(\mathbf{h})$ adopted (Dietrich, 1997). The computational benefit now stems from the structure of the new covariance matrix $\tilde{\mathbf{\Sigma}} = [\sigma_{mm'}, m = 1, \ldots, \tilde{M}, m' = 1, \ldots, \tilde{M}]$, due to the toroidal geometry of the grid and the stationary covariance model $\sigma(\mathbf{h})$. More precisely, matrix $\tilde{\mathbf{\Sigma}}$ is a circulant matrix in 1D, and a block circulant matrix with circulant blocks in higher dimensions (Davis, 1994). A circulant matrix is completely characterized by its first row or column in the 1D case; in higher dimensions one considers the first block row or column. Storage and computation become very efficient, since they only involve the $\tilde{M}$ entries of the first row or column of $\tilde{\mathbf{\Sigma}}$ instead of all its $(\tilde{M} \times \tilde{M})$ entries.

For a regular 1D grid of M nodes with unit spacing, for example, the new $(\tilde{M} \times \tilde{M})$ embedded covariance matrix $\tilde{\mathbf{\Sigma}}$, with $\tilde{M} = 2M$, becomes

$$\tilde{\mathbf{\Sigma}} = \begin{bmatrix} \sigma(0) & \sigma(1) & \cdots & \sigma(M-1) & \sigma(M) & \sigma(M-1) & \cdots & \sigma(1) \\ \sigma(1) & \sigma(0) & \cdots & \sigma(M-2) & \sigma(M-1) & \sigma(M) & \cdots & \sigma(2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma(2) & \sigma(1) & \cdots & \sigma(M-3) & \sigma(M-2) & \sigma(M-1) & \cdots & \sigma(1) \\ \sigma(1) & \sigma(2) & \cdots & \sigma(M-2) & \sigma(M-1) & \sigma(M-2) & \cdots & \sigma(0) \end{bmatrix}$$

which is again completely characterized by its first $(1 \times \tilde{M})$ row vector $[\sigma(0) \cdots \sigma(M) \cdots \sigma(1)]$. The $(j+1)$-th row of $\tilde{\mathbf{\Sigma}}$ is constructed by shifting the preceding j-th row one position to the right, and filling the leading empty position in the $(j+1)$-th row with the last element of the j-th row.

Most importantly, the square root matrix $\sqrt{\tilde{\mathbf{\Sigma}}}$ of a $(\tilde{M} \times \tilde{M})$ circulant covariance matrix $\tilde{\mathbf{\Sigma}}$ can be computed as (Davis, 1994)

$$\sqrt{\tilde{\mathbf{\Sigma}}} = \tilde{\mathbf{F}}\sqrt{\tilde{\mathbf{E}}}\tilde{\mathbf{F}}^H \tag{4.3}$$

where $\tilde{\mathbf{F}}$ denotes the $(\tilde{M} \times \tilde{M})$ unitary and Hermitian Fourier matrix, $\tilde{\mathbf{E}} = diag(\tilde{\mathbf{e}})$ denotes the $(\tilde{M} \times \tilde{M})$ diagonal matrix of eigenvalues of $\tilde{\mathbf{\Sigma}}$ stored in the $(\tilde{M} \times 1)$ vector $\tilde{\mathbf{e}}$, and $\tilde{\mathbf{F}}^H = \tilde{\mathbf{F}}^{-1}$ denotes the $(\tilde{M} \times \tilde{M})$ inverse Fourier matrix, which is the Hermitian transpose of $\tilde{\mathbf{F}}$, i.e., the transpose of its complex conjugate. In other words, *the eigenvectors of (block) circulant matrices are actually the columns of the Fourier matrix* $\tilde{\mathbf{F}}$.

The vector $\tilde{\mathbf{e}}$ of eigenvalues of $\tilde{\mathbf{\Sigma}}$ can be computed efficiently from the first column $\tilde{\boldsymbol{\sigma}}_1$ of matrix $\tilde{\mathbf{\Sigma}}$ as (Davis, 1994)

$$\tilde{\mathbf{e}} = \sqrt{\tilde{M}}\mathbf{F}^H\tilde{\boldsymbol{\sigma}}_1 = \sqrt{\tilde{M}}\mathcal{F}\{\tilde{\boldsymbol{\sigma}}_1\} \tag{4.4}$$

where $\mathcal{F}\{\tilde{\boldsymbol{\sigma}}_1\}$ denotes the Discrete Fourier Transform of vector $\tilde{\boldsymbol{\sigma}}_1$ and is computed very efficiently via the Fast Fourier Transform (FFT) without storing all the entries of matrix $\tilde{\mathbf{F}}^H$.

4.1.3 Simulation on large regular grids

Returning to the simulation or sampling objective, one could to use Eq. (4.2) for generating a realization (at the M nodes of a regular grid) from a random vector with a zero-mean multivariate Gaussian distribution $\mathbf{y} \sim \mathcal{G}(\mathbf{0}, \mathbf{\Sigma})$. Instead, one resorts to the more computationally efficient simulation of a random vector at the $\tilde{M}$ nodes of an extended regular grid with toroidal geometry. The new random vector $\tilde{\mathbf{y}}$ follows a zero-mean multivariate Gaussian distribution $\tilde{\mathbf{y}} \sim \mathcal{G}\left(\tilde{\mathbf{0}}, \tilde{\mathbf{\Sigma}}\right)$, with a $(\tilde{M} \times \tilde{M})$ circulant covariance matrix $\tilde{\mathbf{\Sigma}}$.

A $(\tilde{M} \times 1)$ vector $\tilde{\mathbf{y}}$ with simulated Gaussian deviates at $\tilde{M}$ locations, can be generated using Eq. (4.2), as

$$\tilde{\mathbf{y}} = \sqrt{\tilde{\mathbf{\Sigma}}}\tilde{\mathbf{w}} = \tilde{\mathbf{Q}}\left(\sqrt{\tilde{\mathbf{e}}} \odot (\tilde{\mathbf{Q}}^T\tilde{\mathbf{w}})\right)$$

When those $\tilde{M}$ locations coincide with the nodes of an extended regular grid with toroidal geometry, and the covariance model used is stationary, the above equation, hence

Eq. (4.2), can be re-expressed as (Dietrich, 1997):

$$\tilde{\mathbf{y}} = \tilde{\mathbf{F}}\left(\sqrt{\tilde{\mathbf{e}}} \odot (\tilde{\mathbf{F}}^H \tilde{\mathbf{w}})\right) = \mathcal{F}^{-1}\left\{\sqrt{\sqrt{\tilde{M}}\mathcal{F}\{\tilde{\boldsymbol{\sigma}}_1\}} \odot \mathcal{F}\{\tilde{\mathbf{w}}\}\right\} \tag{4.5}$$

where $\mathcal{F}^{-1}\{\mathbf{v}\}$ denotes the inverse DFT (IDFT) of an arbitrary vector $\mathbf{v}$, which is again computed very efficiently via the inverse FFT. The result is a $(\tilde{M} \times 1)$ vector $\tilde{\mathbf{y}}$ with simulated Gaussian deviates at the nodes of the extended grid, from which one extracts the $(M \times 1)$ vector $\mathbf{y}$ of simulated values at the original M grid nodes; see, Fig. (4.1.

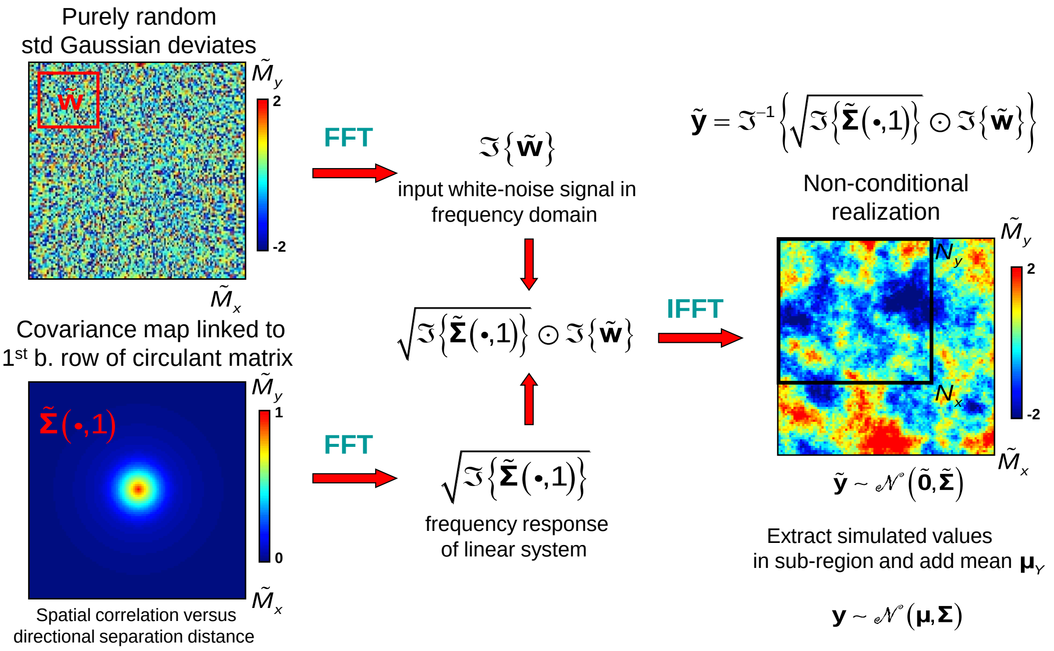

Purely random
std Gaussian deviates
$\tilde{\mathbf{w}}$
$\tilde{M}_y$
2
-2
$\tilde{M}_x$
Covariance map linked to
1st b. row of circulant matrix
$\tilde{\boldsymbol{\Sigma}}(\bullet,1)$
$\tilde{M}_y$
1
0
$\tilde{M}_x$
Spatial correlation versus
directional separation distance
FFT
$\Im\{\tilde{\mathbf{w}}\}$
input white-noise signal in
frequency domain
$\sqrt{\Im\{\tilde{\boldsymbol{\Sigma}}(\bullet,1)\}} \odot \Im\{\tilde{\mathbf{w}}\}$
FFT
$\sqrt{\Im\{\tilde{\boldsymbol{\Sigma}}(\bullet,1)\}}$
frequency response
of linear system
IFFT
$\tilde{\mathbf{y}} = \Im^{-1}\left\{\sqrt{\Im\{\tilde{\boldsymbol{\Sigma}}(\bullet,1)\}} \odot \Im\{\tilde{\mathbf{w}}\}\right\}$
Non-conditional
realization
$\tilde{M}_y$
N_y
N_x
2
-2
$\tilde{M}_x$
$\tilde{\mathbf{y}} \sim \mathcal{N}(\tilde{\mathbf{0}}, \tilde{\boldsymbol{\Sigma}})$
Extract simulated values
in sub-region and add mean $\boldsymbol{\mu}_Y$
$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Figure 4.1: Schematic example illustrating the generation of one realization from a Gaussian random field model on a large grid; see text for details; the figure is not drawn to scale

Simulation flowchart: In summary, our proposed method for Latin hypercube sampling of a Gaussian random field with a stationary covariance model $\sigma(\mathbf{h})$ proceeds in the following steps:

1. Compute and store only the first (block) column $\tilde{\boldsymbol{\sigma}}_1$ of the $(\tilde{M} \times \tilde{M})$ embedded covariance matrix $\tilde{\boldsymbol{\Sigma}}$
2. Compute the square root of the $\tilde{M}$ eigenvalues of matrix $\tilde{\boldsymbol{\Sigma}}$ via FFT of $\tilde{\boldsymbol{\sigma}}_1$ as: $\sqrt{\tilde{\mathbf{e}}} = \sqrt{\sqrt{\tilde{M}}\mathcal{F}\{\tilde{\boldsymbol{\sigma}}_1\}}$ – Eq. (4.4)
3. Loop over the S realizations of the Gaussian random field and:
   (a) generate a $(\tilde{M} \times 1)$ vector $\tilde{\mathbf{w}}$ of uncorrelated standard Gaussian deviates and compute its FFT $\mathcal{F}\{\tilde{\mathbf{w}}\}$ – see Eq. (4.5)
   (b) compute the Hadamard product of $\sqrt{\tilde{\mathbf{e}}}$ and $\mathcal{F}\{\tilde{\mathbf{w}}\}$ as: $\sqrt{\sqrt{\tilde{M}}\mathcal{F}\{\tilde{\boldsymbol{\sigma}}_1\}} \odot \mathcal{F}\{\tilde{\mathbf{w}}\}$, and compute the inverse FFT of the result – see Eq. (4.5). This step yields a $(\tilde{M} \times 1)$ vector $\tilde{\mathbf{y}}$ of zero-mean Gaussian deviates at the $\tilde{M}$ nodes of the extended grid; the spatial correlation of those values is controlled by the covariance function $\sigma(\mathbf{h})$
   (c) extract the $(M \times 1)$ vector $\mathbf{y}$ of simulated Gaussian deviates at the original M simulation grid nodes, and add the $(M \times 1)$ expectation (mean) vector $\boldsymbol{\mu}$

   Repeating the above steps S times yields a SR sample of size S from the Gaussian random field model; these S realizations can be stored into a $(S \times M)$ matrix $\mathbf{Y}_R$, whose covariance approximates $\boldsymbol{\Sigma}$
4. Transform the above SR sample of matrix $\mathbf{Y}_R$ into a LH sample, a $(S \times M)$ matrix $\mathbf{Y}_L$, using Stein's method in Eq. (2.24); the entries of the M columns of matrix $\mathbf{Y}_L$ are marginally stratified into S strata, and are pairwise correlated with the entries of any other column of $\mathbf{Y}_L$

The result of the above procedure is a LH sample of size S from a Gaussian random field on a possibly very large (with millions of nodes) regular grid.

Two realizations (out $S = 10$ generated) from a second-order stationary Gaussian random field on a 2D regular grid of 1000×1000 nodes with unit node spacing are shown in Fig. (4.2). The covariance model specified is of exponential form with a unit sill and an effective range of 100 distance units.

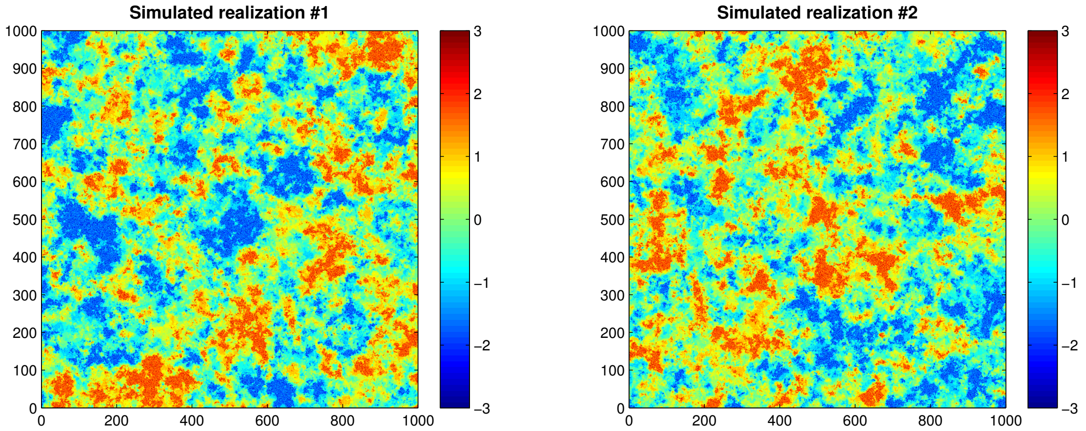


Figure 4.2: Realizations from Gaussian random field models with isotropic (left) and anisotropic (right) covariance function, on a 1000 × 1000 grid with unit spacing.

Last, Fig. (4.3) shows scatter plots of simulated values at two randomly selected pairs of grid nodes verifying the marginal stratification of LH sampling in both cases.

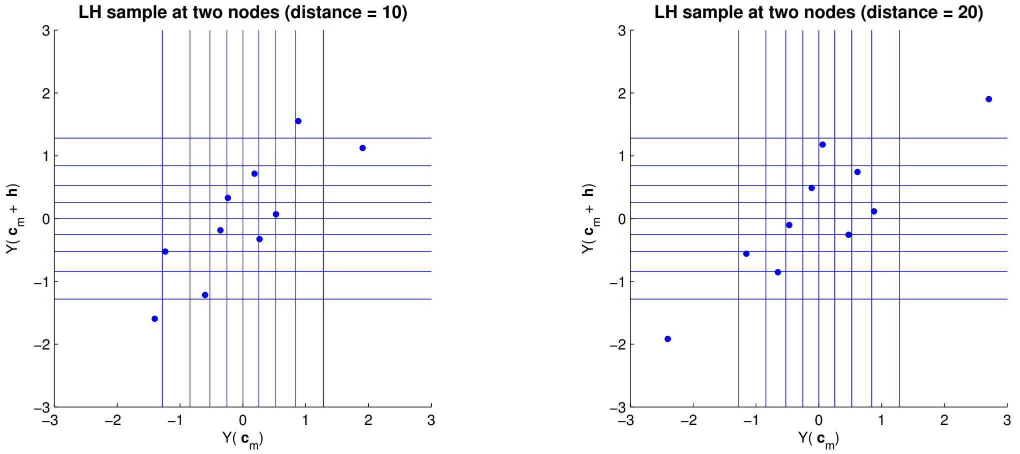


Figure 4.3: Two scatter plots of simulated values extracted from two randomly selected pairs of nodes separated by distance 10 (left) and 20 (right). Vertical and horizontal lines correspond to deciles of a std Gaussian RV and delineate strata of equal probability; see text for details.

4.2 Synthetic case studies

This Section presents two case studies (2D and 3D) comparing simple random and Latin hypercube sampling from a lognormal random field modeling the spatial distribution of saturated hydraulic conductivity in hydrogeological context involving flow and transport in a partially heterogeneous porous medium. The comparison between the two methods is performed along the lines of Sec. 2.4. More precisely, the two sampling methods are firstly evaluated in terms of their ability to reproduce ensemble statistics of the random field; these statistics are computed from a very large set of conductivity realizations generated via simple random sampling. That large set of conductivity fields is used to derive a large set of solute concentration fields through flow and transport simulation. The two sampling methods are then compared again in terms of their ability to reproduce ensemble statistics of the solute concentration field.

4.2.1 2D Hydraulic conductivity

A two-dimensional synthetic groundwater flow system is considered, similar to that used in (Zhang and Pinder, 2003). The dimensions of the flow system are 5005 m by 5005 m discretized into a 1001 × 1001 grid with uniform rectangular cells of size 5 m by 5 m. Porosity was assumed constant throughout the domain and equal to 0.25, and the parametrization of hydraulic conductivity is described below. Flow boundary conditions consisted of constant head of 0 m at the four corner cells and a constant head of 250 m at the central cell of the domain. No flow conditions ($\partial h/\partial n = 0$) were assigned to the rest of the domain boundaries, whereas the Modflow code of McDonald and Harbaugh (1988) was used in this study to obtain the steady state flow solution.

A second-order stationary and isotropic lognormal random field is adopted with parameters borrowed from (Sudicky et al., 2010). More precisely, the mean and variance of log conductivity are taken as $\mu_Y = -5.64$ and $\sigma_Y^2 = 1.79$, respectively, corresponding to conductivity statistics $\mu_Z = 0.0087$ m/sec and $\sigma_Z^2 = 0.0194^2$ (m/sec)^2. The semivariogram of log conductivity is assumed to be of exponential form, with no nugget effect, and effective range 1001 m, corresponding to one fifth of the domain extent along the cardinal directions. Two realizations of this random field model at the nodes of the grid are given in Fig. 4.4.

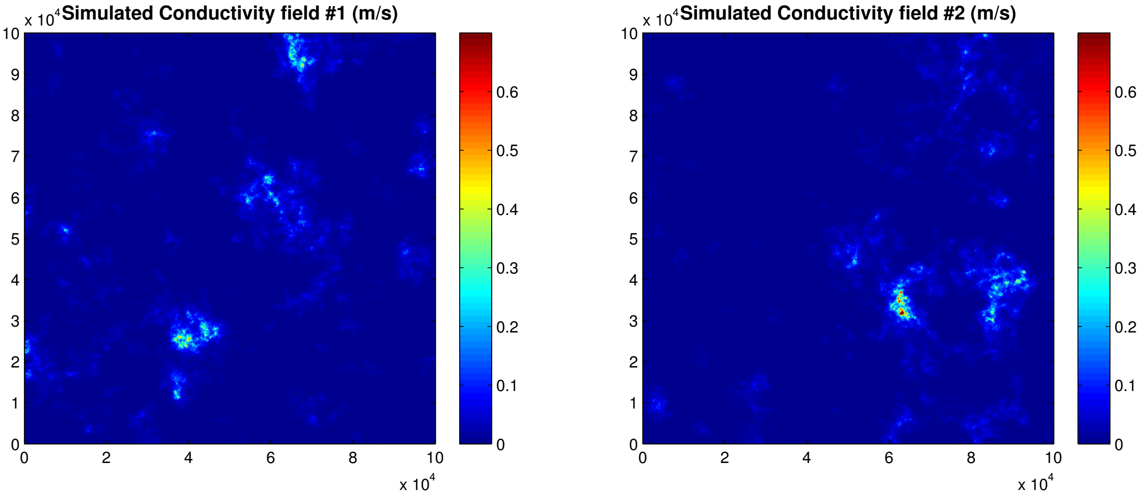


Figure 4.4: Two realizations of a lognormal random field; see text for details.

Reference ensemble statistics are derived from a set of 1000 realizations of hydraulic conductivity generated via simple random (SR) sampling, and consist of: (i) the ensemble average field, – a constant equal to $\mu_Z = 0.0087$ and (ii) the ensemble standard deviation field.

Two sampling or simulation methods are considered in this case study for generating realizations of a lognormal hydraulic conductivity field. These methods include simple random (SR) sampling and Latin hypercube (LH). In terms of sample size or number of realizations per method, three such sizes are considered; namely, $S = 20, 50$, and 80. Once a sample, say of size $S = 20$, is generated, the discrepancy between the statistics of the simulated ensemble and the reference conductivity statistics listed above is quantified using the root mean squared error (RMSE) for summary statistics (mean, variance). The computation of such error statistics is repeated over a set of $I = 100$ batches of realizations, with each batch containing the same sample size, $S = 20$ for example, thus estimating the sampling distributions of RMSE values for each sample size and for each method; these distributions are presented in terms of their means and medians, as well as their 75% and 95% probability intervals. The better the reproduction of a reference statistic from simulations of a sampling method with a given sample size, the narrower the sampling distribution of the resulting RMSE values and the smaller (closer to 0) the center of that distribution.

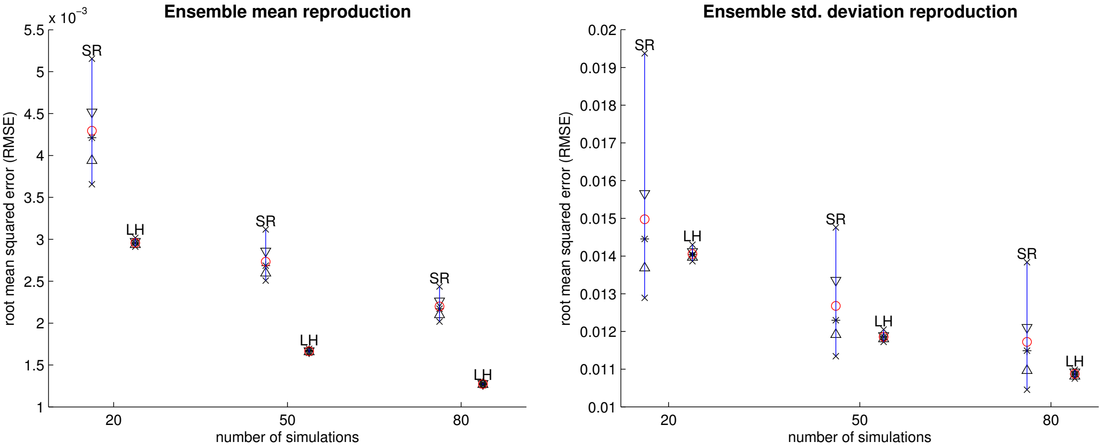


Figure 4.5: Reproduction of reference ensemble mean (A) and ensemble standard deviation (B) hydraulic conductivity fields from the two sampling methods; see text for details.

More precisely, the reproduction of the reference ensemble mean and ensemble standard deviation of hydraulic conductivity from the two sampling methods and the three sample sizes considered is shown in Fig. 4.5. Reproduction is quantified in terms of the sampling distribution of RMSE between reference and simulated ensemble conductivity statistics over the 1001 × 1001 grid cells. In this and all subsequent figures, 75% RMSE (or MAE depending on the statistic selected) probability intervals are depicted with horizontal line segments, whereas 95% probability intervals with × symbols; median RMSE values are depicted as asterisks ($*$), whereas mean values as circles ($\circ$). From Fig. 4.5, it can be readily appreciated that LH sampling (LH) yields the closest reproduction for both the ensemble mean and standard deviation statistics. This is expected, since LH sampling aims at marginal stratification, hence should best reproduce marginal hydraulic conductivity statistics at each grid node.

4.2.2 2D Solute concentration

For the solute transport problem, an initial concentration equal to 0 is assumed throughout the model domain. At time $t = 0$, a contaminant is introduced at the central cell, along the upstream constant head boundary, with constant concentration $C_0 = 100$ mg/l. No transport conditions ($\partial C/\partial n = 0$) are assigned along the domain boundaries. Longitudinal and transverse horizontal dispersivities are assumed to be equal to 5 m and 0.5 m, respectively. In terms of software, the MT3D code Zheng (1990) was used to obtain breakthrough curves at the four observation wells located at the four corner nodes of the domain, as well as the solute transport solution up to time $t = 5 \cdot 10^6$ sec. Reference ensemble statistics for solute concentration are derived from a set of 1000 solutions of the transport problem based on the 1000 hydraulic conductivity realizations generated in the previous subsection via SR sampling. Two such concentration realizations derived from the conductivity realizations shown in Fig. 4.4 are given in Fig. 4.6.

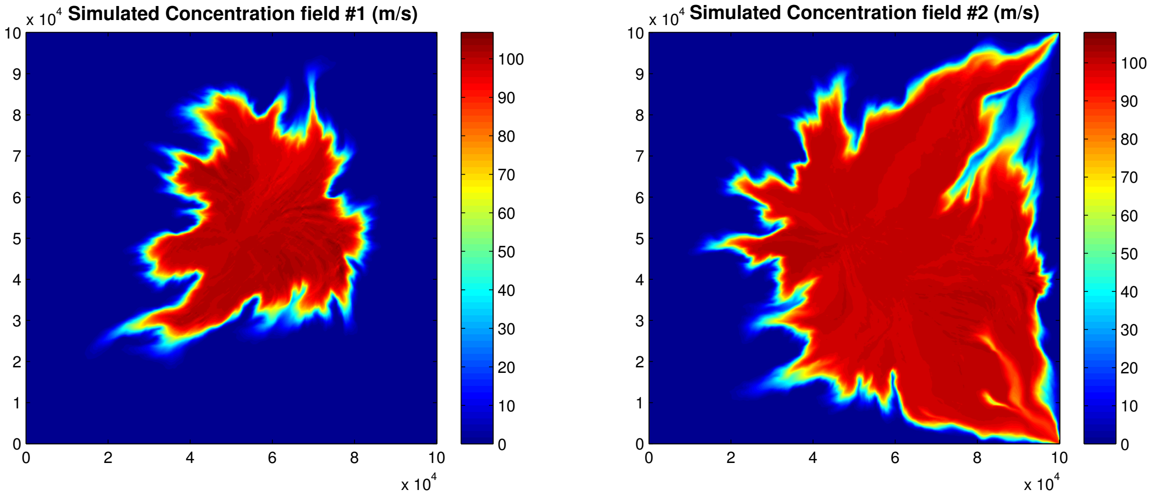


Figure 4.6: Two solute concentration realizations corresponding to the two hydraulic conductivity realizations shown in Fig. 4.4; see text for details.

The performance of the two sampling methods considered in this work for sample sizes $S = 20, 50, 80$ is then quantified in terms of reproduction of these reference ensemble concentration statistics via the sampling distribution of error summary statistics (RMSE or MAE). Reference statistics for solute concentration consist of: (i) the ensemble average field shown in Fig. 4.7, (ii) the ensemble standard deviation field also shown in Fig. 4.7 and (iv) the ensemble distribution of first arrival times (Fig. 4.7) at any of the four observation wells.

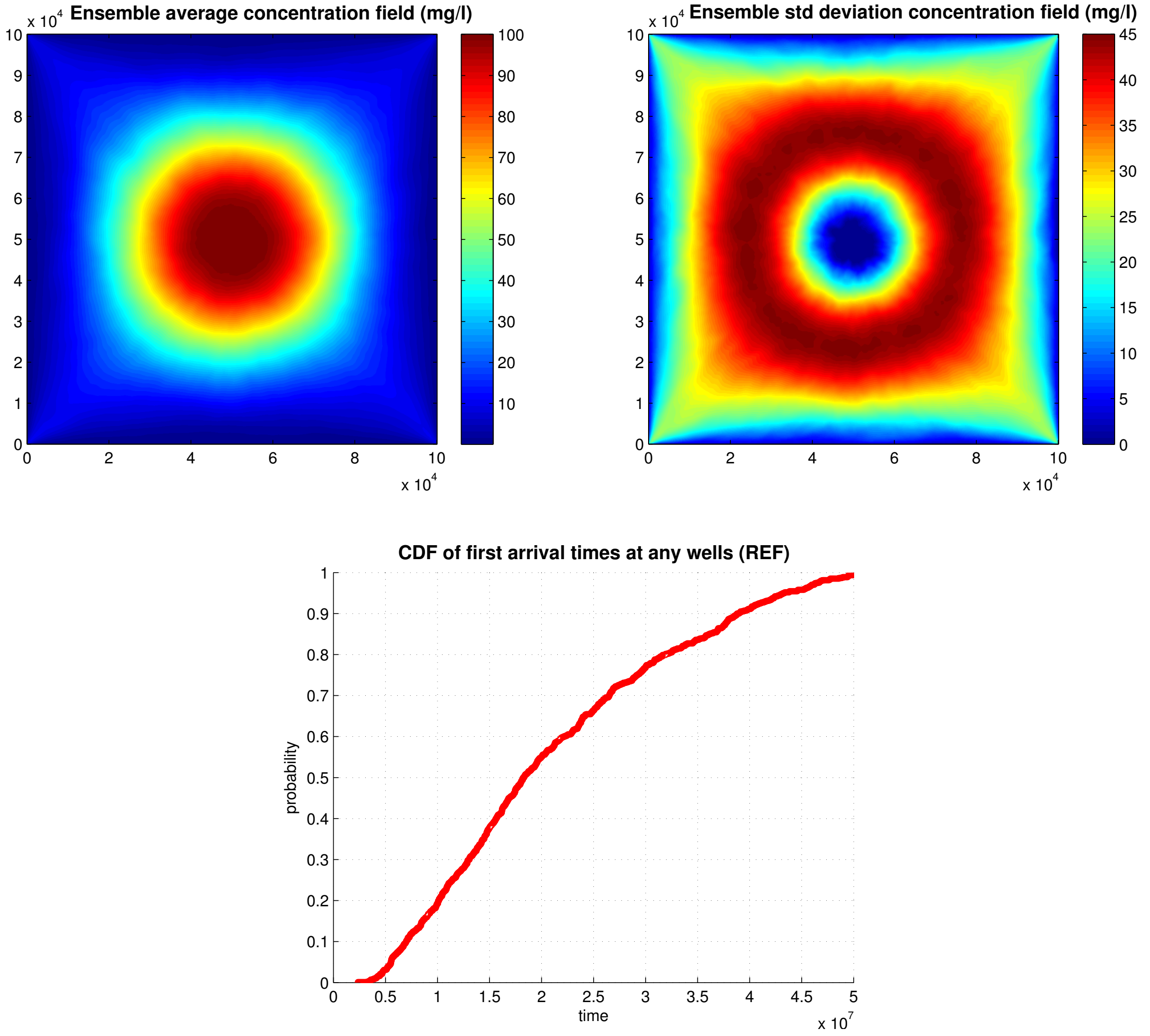


Figure 4.7: Reference ensemble average (top left), ensemble standard deviation (top right) concentration fields, and ensemble first arrival time CDF (bottom), computed from 1000 concentration realizations derived from 1000 hydraulic conductivity realizations generated via simple random sampling; see text for details.

Figure 4.8 shows the reproduction of the ensemble average and ensemble standard deviation concentration fields (Fig. 4.7) for the two sampling methods and the three sample sizes considered. Reproduction is quantified here in terms of the sampling distribution of RMSE between reference and simulated ensemble concentration statistics over the 1001×1001 grid cells. It is easily appreciated that LH yields a better reproduction of these ensemble fields than SR sampling, particularly in the case of the ensemble average concentration field.

Last, Fig. 4.9 gives the reproduction of the reference cumulative distribution function (CDF) of first arrival times at any of the four observation wells at the corners of the domain (Fig. 4.7 bottom), and the corresponding CDFs obtained from the two sampling methods and the three sample sizes considered. One can easily identify the shorter range

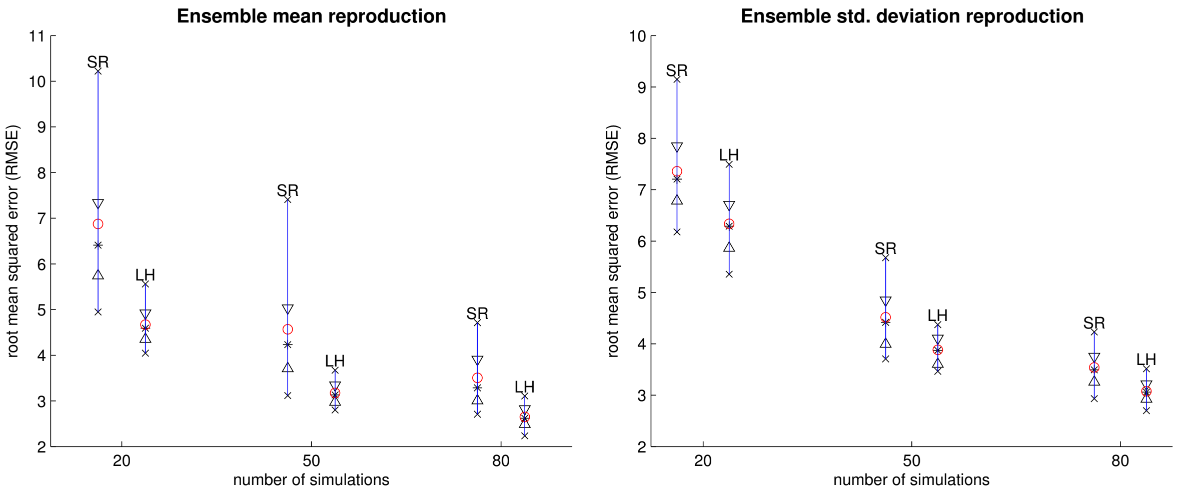


Figure 4.8: Reproduction of reference ensemble mean (A) and ensemble standard deviation (B) concentration fields (shown in Fig. 4.7) from two sampling methods; see text for details.

of the 100 CDFs corresponding to the sampling method of LH for the three different sample sizes $20-50-80$ (titled as Stein at 4.9B), compared to the CDFs corresponding to the SR sampling method (titled as Simple Random at 4.9A). The more efficient reproductive ability of the stratified method is more distinct at Fig. 4.10, where the discrepancy between any two CDFs is quantified in terms of the mean absolute error (MAE) between their corresponding percentile. The sampling distribution of this statistic is established over $I = 100$ batches of realizations as before.

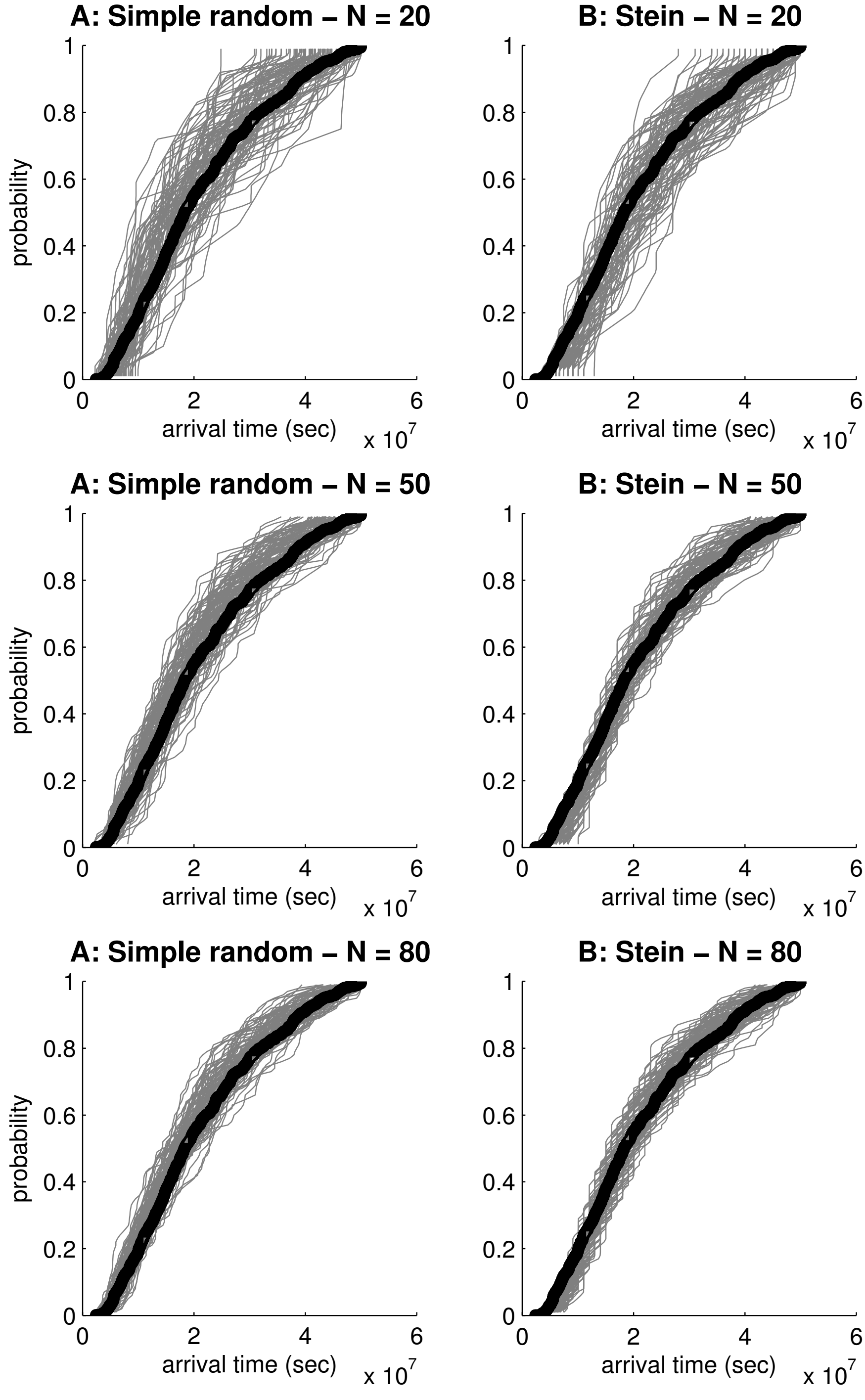


Figure 4.9: Reproduction of reference first arrival time cumulative distribution function (CDF) derived for two sampling methods; see text for details.

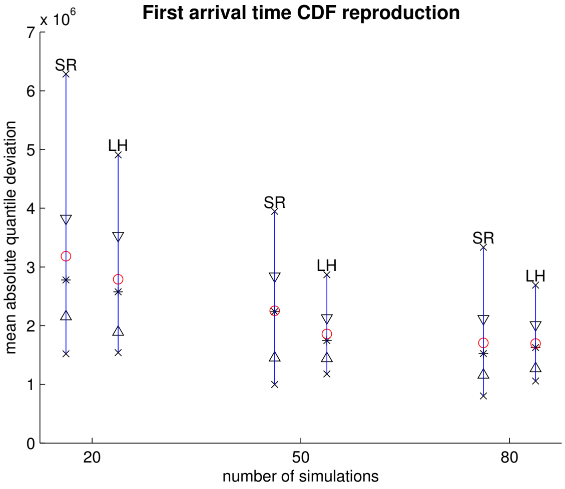


Figure 4.10: Distribution of reproduction of reference first arrival time cumulative distribution function (CDF), derived for two sampling methods; see text for details.

4.2.3 3D Hydraulic conductivity

The following two subsections describe a synthetic case study involving a three dimensional groundwater flow system, discretized using a very large grid (> 1.000.000 nodes). Simple random (SR) and Latin hypercube (LH) sampling are also considered in this case study for generating realizations from a lognormal random field, modelling the spatial distribution of saturated hydraulic conductivity in a hydrogeological flow and transport problem.

More precisely, the dimensions of the flow system are 1005 m by 1005 m by 125 m, discretized into a 201 × 201 × 25 grid with uniform cells of size 5 m by 5 m. Porosity is assumed constant throughout the domain and equal to 0.25. No flow conditions ($\partial h/\partial n = 0$) are assigned to the rest of the domain boundaries, and the Modflow software of McDonald and Harbaugh (1988) is used to obtain the steady state flow solution. For the solute transport problem, an initial concentration equal to 0 mg/l is assumed throughout the model domain. At time $t = 0$, a contaminant is introduced at the central cell (middle of three dimensional domain), along the upstream constant head boundary, with constant concentration $C_0 = 100$ mg/l, and the MT3D software Zheng (1990) is used to solve the transport problem. Fig. 4.11 illustrates a pictorial representation of a three dimensional simulated conductivity field (left) and the respective concentration realization (right).

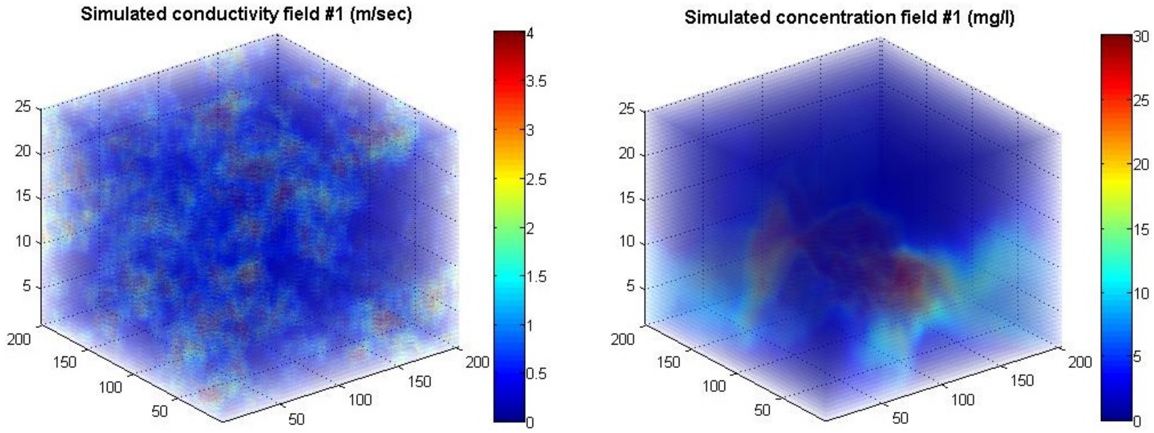


Figure 4.11: A realization of a 3D lognormal random field (left), and corresponding 3D solute concentration realization (right).

Reference ensemble statistics are derived from a set of 1000 realizations of hydraulic conductivity generated via SR sampling and consist of: (i) the ensemble average field – a constant equal to 0.0087 m/sec, and (ii) the ensemble standard deviation field – a constant equal to 0.0194 m/sec.

In terms of sample size or number of realizations per method, three such sizes are considered; namely, $S = 20, 50$, and 80. Once a sample, say of size $S = 20$, is generated, the discrepancy between the statistics of the simulated ensemble and the reference statistics is quantified using the root mean squared error (RMSE). The computation of such error statistic is repeated over a set of $I = 100$ batches of realizations, with each batch containing the same sample size, $S = 20$ for example, thus estimating the sampling distributions of RMSE values for each sample size and for each method; these distributions are presented in terms of their means and medians, as well as their 75% and 95% probability intervals.

Fig. 4.12 depicts the reproduction of the ensemble average and ensemble standard deviation of hydraulic conductivity for the two sampling methods and the three sample sizes considered. It is easily appreciated from both figures that LH yields a better reproduction of these ensemble average and standard deviation fields than SR sampling.

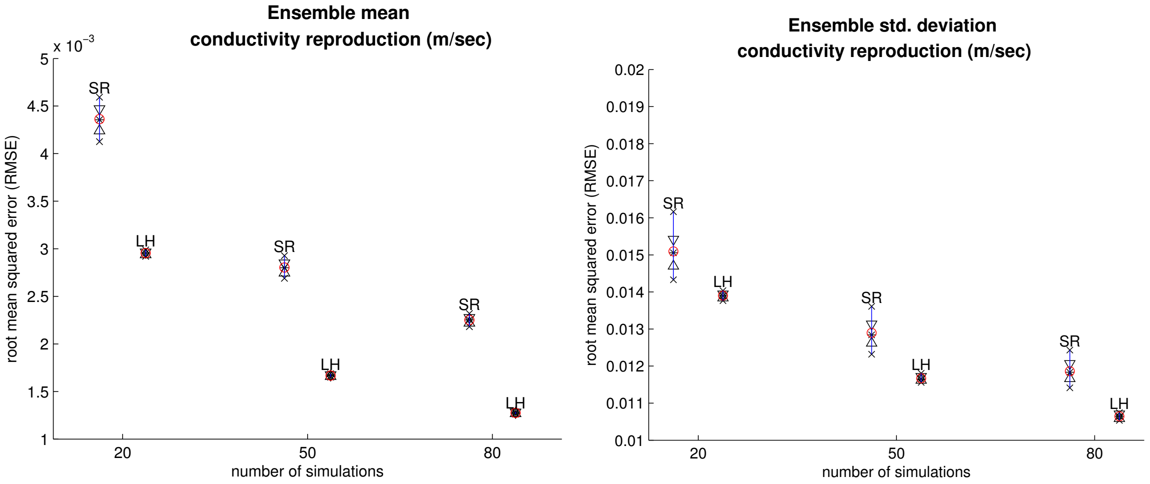


Figure 4.12: Reproduction of reference ensemble mean (left) and standard deviation (right) fields for hydraulic conductivity; see text for details.

4.2.4 3D Solute concentration

Reference ensemble statistics for solute concentration are derived from a set of 1000 solutions of the transport problem based on the 1000 hydraulic conductivity realizations generated in the previous subsection via SR sampling. The reference statistics for solute concentration consist of: (i) the ensemble average field shown in Fig. 4.13 (left), and (ii) the ensemble standard deviation field shown in Fig. 4.13 (right).

Fig. 4.14 depicts the reproduction of the ensemble mean and standard deviation fields of concentration for the two sampling methods and the three sample sizes considered. Here again, LH yields a better reproduction of these ensemble average and standard deviation fields than SR sampling for solute concentration.

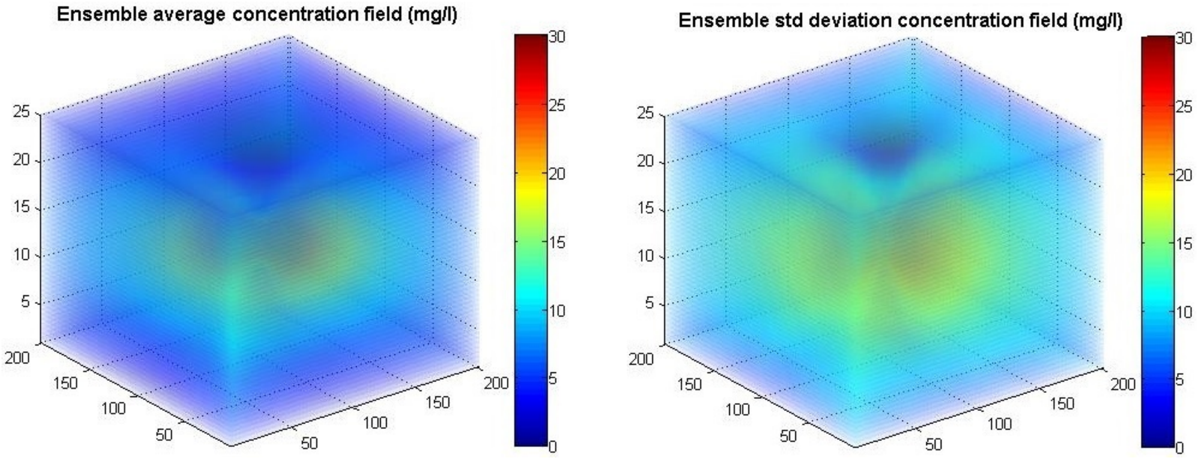

Ensemble average concentration field (mg/l)
Ensemble std deviation concentration field (mg/l)
25
20
15
10
5
200
150
100
50
30
0

Figure 4.13: Reference ensemble average (left) and standard deviation (right) concentration fields, computed from 1000 concentration realizations derived from 1000 hydraulic conductivity realizations generated via simple random sampling.

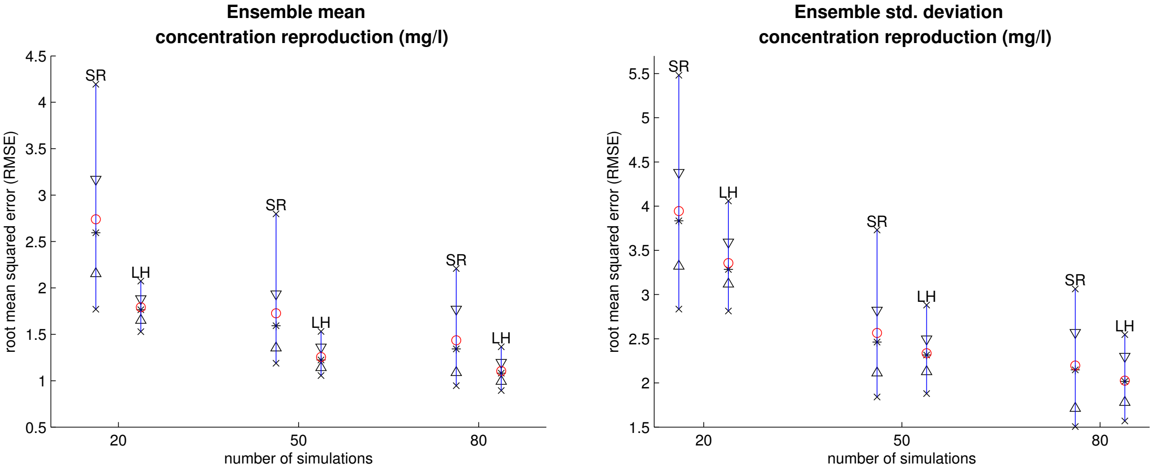


Figure 4.14: Reproduction of reference ensemble mean (left) and standard deviation (right) fields for concentration; see text for details.

4.3 Conclusions and Discussion

In hydrogeological investigations involving flow and transport in heterogeneous porous media, the spatial distribution of saturated hydraulic conductivity is often parameterized in terms of a lognormal random field model. Simulated realizations of such a random field (simulated attribute values arranged on two or three dimensional grids) are then used along with physically-based models of flow and transport in a Monte Carlo framework for evaluating, for example, the uncertainty in the spatial distribution of solute concentration due to the uncertainty in the spatial distribution of hydraulic conductivity and possibly other relevant variables. In practical applications, however, classical Monte Carlo simulation, or simple random (SR) sampling, from a random field model can quickly become computationally expensive as the size of the simulation grid involved is typically very large, particularly for 3D investigations.

A more computationally efficient alternative to classical Monte Carlo simulation is Latin hypercube (LH) sampling, a form of stratified random sampling, which aims at reducing the number of realizations that are too similar by chance. In a spatial context, LH sampling aims at generating representative realizations from a random field model, where term representative implies realizations spanning efficiently the range of possible attribute realizations corresponding to the multivariate probability distribution (lognormal in this work) associated with the random field. In practice, LH sampling amounts to generating S spatially correlated realizations (a sample of size S) of hydraulic conductivity from the random field model, whose S simulated values at each grid node are also (marginally) stratified in S strata. Existing methods for Latin hypercube sampling of random fields, however: (i) require storage of a covariance matrix pertaining to all possible pairs of grid nodes, and (ii) call for the Cholesky decomposition of that covariance matrix. Consequently, such methods cannot handle a large number of simulation locations Chilès and Delfiner (1999).

The proposed solution to the above problem derives from the fact that LH sampling is actually a post-processing method, that transforms correlated SR samples into marginally stratified correlated LH samples; Sec 2.3. More precisely, Stein's LH sampling method uses a SR sample as input at generating the LH stratified sample. Hence, since the proposed LH method does not incorporate the Cholesky decomposition of the covariance matrix, can be applied at very large grids, if a proper method exists for generating SR samples at those grids.

Along these lines, this paper proposes a novel simulation method for generating Latin hypercube samples from second order stationary Gaussian random fields at very large grids (with millions of nodes). The proposed LH sampling method combines an existing simulation method for creating SR samples from second order stationary random fields in large regular grids Dietrich (1997), with the LH sampling method of (Stein, 1987).

The performance of the proposed LH sampling method was compared to that of simple random (SR) sampling via a synthetic case study involving a three dimensional flow and transport system, assuming a porous medium whose saturated hydraulic conductivity is parameterized in terms of a lognormal random field with a stationary covariance model. Performance comparison for the two methods included reproduction of: (a) the reference ensemble average and ensemble standard deviation conductivity, and (b) the reference ensemble average and standard deviation concentration fields derived from concentration realizations obtained by solving a flow and transport boundary value problem for each hydraulic conductivity realization, along with the first arrival times of a non zero pollutant at one of the four observation wells at the corners of the domain. The results indicate that LH sampling exhibits a much smaller sampling variability, than SR sampling, for the same number of simulated realizations, both for the case of hydraulic conductivity and solute concentration. In this case, LH facilitates efficient uncertainty propagation with fewer model runs due to more representative model inputs. It could be thus argued that the proposed LH sampling method could reduce the time and computer resources required to perform uncertainty analysis in hydrogeological flow and transport problems discretized by very large regular grids.

According to an extensive study of the literature that was carried out, it is the first time the stratified LH sampling method is applied for the simulation of random fields in large grids. Hence, it is the first time that efficiently geostatistical simulation can be preformed at models with detailed discretization of the spatially distributed parameters.

Chapter 5

Conditional Latin Hypercube Sampling

The propagation of the inherent spatial variability of petrophysical attributes, such as hydraulic conductivity, to the predictions of physically-based, numerical models simulating flow and transport is a critical requirement in hydrogeological investigations. In a classical, two-point, geostatistical context, that spatial attribute variability is typically modelled via a random field, log-normal in the case of hydraulic conductivity, which is often parameterized by a spatially constant mean and a variogram model. Conditional realizations from such a random field generated via geostatistical simulation using simple random (SR) sampling are employed, along with the numerical simulators of flow and transport, in a Monte Carlo framework for evaluating, for example, the uncertainty in the spatial distribution of solute concentration. The number of model runs, however, required to furnish a representative distribution of model outputs, often renders classical Monte Carlo simulation based on SR sampling computationally challenging, thus hindering the widespread application of uncertainty and/or sensitivity analyses in practice.

In this chapter, a novel conditional simulation method based on Latin hypercube (LH) sampling is presented, a form of stratified random sampling, for generating, very fast, marginally (per node) stratified conditional realizations of stationary Gaussian (or transformed Gaussian) random fields. Contrary to existing methods of spatial LH sampling that rely on the Cholesky decomposition of the covariance matrix, the proposed method can furnish realizations on very large (comprised of millions of nodes) three dimensional (in the implemented case study) regular grids. The proposed LH-based simulation method is employed for generating three dimensional realizations of hydraulic conductivity in a synthetic case study involving flow and transport in a mildly heterogeneous porous medium, whereby the classical (two-point) geostatistical framework is typically adopted. Even when the additional uncertainty in the parameters (mean and variogram model) of the random field are accounted for, the results indicate that the proposed conditional LH simulation method can reproduce statistics of the conductivity and concentration fields with smaller sampling variability than SR sampling for the same number of realizations. Viewed from a computational perspective, LH sampling requires approximately one third of the time needed by SR sampling to reach model results of equal reliability (sampling error).

In this Chapter, a novel method is developed for conditional Latin hypercube (LH) sampling of second-order stationary Gaussian random fields on large regular grids. Section 5.1 describes in detail the theory of conditional sampling and provides two alternative methods of implying it, Section 5.5 describes the combination of LH sampling of random fields on very large grids, and conditional simulation via indirect simulation of the Kriging error, and Section 5.6 presents a synthetic case study involving flow and transport in a heterogeneous porous medium illustrating the benefits of the proposed conditional Latin hypercube over simple random sampling.

5.1 Introduction

This section describes a conditional sampling method for generating S alternative joint realizations of y-values at all M target supports, given (conditioned on) the sample data vector $\mathbf{d}$ available at the N source supports and the known expectations; in statistical jargon: generate S samples from conditional multivariate distribution of M (target) RVs given realizations or samples of N (source) RVs. The generated S realization is reproducing (i) sample data at their supports with no artifact discontinuities, (ii) data variogram, or a model for it, and (iii) data histogram, or a model for it. Fig. 5.1 holds the unknown attribute values (left) on the M target supports and the known source attribute values (right) on the N source supports.

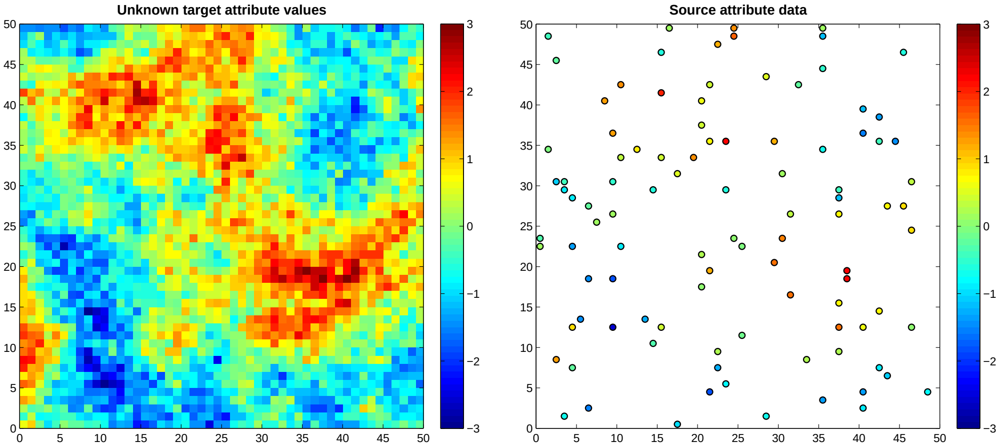


Figure 5.1: Example of target attribute values (left) and source data (right)

Fig. 5.2 illustrates two generated conditional realizations from the source attribute values and the histogram and variogram of the target unknown attribute values.

Conditional sampling generates $(M \times 1)$ vector $\mathbf{y} = [y(t_m), m = 1, \ldots, M]^T$ of attribute values at M target supports (these M values are viewed as a realization of a $(M \times 1)$ Gaussian random vector). Applying conditional sampling considers as known (i) the $(M \times 1)$ vector $\boldsymbol{\mu}_Y = [\mu_Y(t_m), m = 1, \ldots, M]^T = \mathbb{E}\{\mathbf{y}\}$ of prior (to any measurement acquisition) attribute expectations at M target supports, (ii) the $(N \times 1)$ data vector $\mathbf{d} = [d(s_n), n = 1, \ldots, N]^T$ of measurements at N source supports, possibly corrupted by

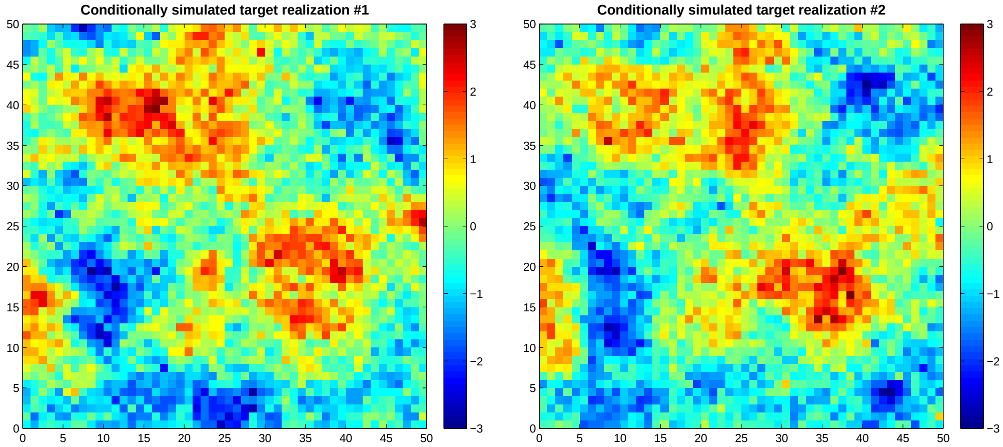


Figure 5.2: Example of simulated values conditioned on reproducing the source data (5.1 right) and the histogram and variogram of the unknown target attribute values (5.1 left)

measurement error; these N data values are regarded as a realization of a $(N\times 1)$ Gaussian random vector, and (iii) the $(N \times 1)$ vector $\boldsymbol{\mu}_D = [\mu_D(s_n), n = 1, \ldots, N]^T = \mathbb{E}\{\mathbf{d}\}$ of prior attribute expectations at N source supports.

Taking into account the above prior information all three combinations of covariances between random variables on source (data) and target locations are calculated as:

- target-to-target covariance matrix:
  $\boldsymbol{\Sigma}_{tt} = [\sigma_Y(t_m, t_{m'}), m = 1, \ldots, M, m' = 1, \ldots, M] = (M \times M)$ matrix of covariance values between all pairs of target RVs:

  $$\boldsymbol{\Sigma}_{YY} = \mathbb{E}\{[\mathbf{y} - \boldsymbol{\mu}_Y][\mathbf{y} - \boldsymbol{\mu}_Y]^T\} = \mathbb{C}ov\{\mathbf{y}\}$$

- target-to-source covariance matrix:
  $\boldsymbol{\Sigma}_{YD} = [\sigma_{YD}(t_m, s_n), m = 1, \ldots, M, n = 1, \ldots, N] = (M \times N)$ matrix of covariance values between all pairs of target and source RVs:

  $$\boldsymbol{\Sigma}_{YD} = \mathbb{E}\{[\mathbf{y} - \boldsymbol{\mu}_Y][\mathbf{d} - \boldsymbol{\mu}_D]^T\} = \mathbb{C}ov\{\mathbf{y}, \mathbf{d}\}$$

  note that: $\boldsymbol{\Sigma}_{DY} = \mathbb{E}\{[\mathbf{d} - \boldsymbol{\mu}_D][\mathbf{y} - \boldsymbol{\mu}_Y]^T\} = \boldsymbol{\Sigma}_{YD}^T$

- source-to-source covariance matrix:
  $\boldsymbol{\Sigma}_{ss} = [\sigma_D(s_n, s_{n'}), n = 1, \ldots, N, n' = 1, \ldots, N] = (N \times N)$ matrix of covariance values between all pairs of source RVs:

  $$\boldsymbol{\Sigma}_{DD} = \mathbb{E}\{[\mathbf{d} - \boldsymbol{\mu}_D][\mathbf{d} - \boldsymbol{\mu}_D]^T\} = \mathbb{C}ov\{\mathbf{d}\}$$

  For additive measurement error $\mathbf{d} = \mathbf{y}_s + \mathbf{e}$, independent of the error-free source values $\mathbf{y}_s$, $\boldsymbol{\Sigma}_{DD} = \boldsymbol{\Sigma}_{YY}^{ss} + \boldsymbol{\Sigma}_{EE}$, where $\boldsymbol{\Sigma}_{EE}$ is a $(N \times N)$ matrix of measurement error (co)variances

In what follows, we assume second-order stationarity, i.e.: $\boldsymbol{\mu}_Y = \mu_Y \mathbf{1}_m$, and $\boldsymbol{\mu}_D = \mu_Y \mathbf{1}_n$; $\mu_Y =$ stationary mean

5.2 Simple Kriging

Kriging is a group of geostatistical techniques to interpolate the value of a random field at an unobserved location from observations of its value at nearby locations. The main idea of kriging is that near sample points should get more weight in the prediction to improve the estimate. Thus, kriging relies on the knowledge of some kind of spatial structure, which is modeled via the variogram of the underlying random function. Depending on the stochastic properties of the random field and the various degrees of stationarity assumed, different methods for calculating the weights can be deduced, i.e. different types of kriging alternative methods; some of the classical methods are Ordinary Kriging, Simple Kriging, Universal Kriging, and Indicator Krigig. For a more detailed reference Kriging methods the reader is referred to (Goovaerts, 1997).

Simple Kriging predictions $\hat{\mathbf{y}} = [\hat{y}(t_m), m = 1, \ldots, M]^T$ on the M target supports is an $(M \times 1)$ vector of interpolated values expressed in matrix terms as:

$$[\hat{\mathbf{y}} - \boldsymbol{\mu}_Y] = \boldsymbol{\Lambda}^T[\mathbf{d} - \boldsymbol{\mu}_D] \tag{5.1}$$

where $\boldsymbol{\Lambda}^T = [w_m(\mathbf{s}_n), m = 1, \ldots, M, n = 1, \ldots, N]$ is an $(M \times N)$ matrix of weights:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1(s_1) & \cdots\lambda_m(s_1)\cdots & \lambda_M(s_1) \\ \vdots & \vdots & \vdots \\ \lambda_1(s_N) & \cdots\lambda_m(s_N)\cdots & \lambda_M(s_N) \end{bmatrix} = [\lambda_1 \cdots \lambda_m \cdots \lambda_M]$$

and has as many columns (M) as target supports. Its m-th row is the $(1 \times N)$ vector of weights λ_m^T assigned to N source data for prediction at m-th target support. Simple Kriging system is equal to

$$\boldsymbol{\Sigma}_{DD}\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_{DY} \quad \text{or} \quad \boldsymbol{\Sigma}_{DD}[\lambda_1 \cdots \lambda_m \cdots \lambda_M] = [\boldsymbol{\sigma}_1 \cdots \boldsymbol{\sigma}_m \cdots \boldsymbol{\sigma}_M]$$

where $\boldsymbol{\Sigma}_{DY}$ is

$$\boldsymbol{\Sigma}_{DY} = \begin{bmatrix} \sigma_{DY}(s_1, t_1) & \cdots\sigma_{DY}(s_1, t_m)\cdots & \sigma_{DY}(s_1, t_M) \\ \vdots & \vdots & \vdots \\ \sigma_{DY}(s_N, t_1) & \cdots\sigma_{DY}(s_N, t_m)\cdots & \sigma_{DY}(s_N, t_M) \end{bmatrix} = [\boldsymbol{\sigma}_1 \cdots \boldsymbol{\sigma}_m \cdots \boldsymbol{\sigma}_M]$$

and has as many columns (M) as target supports; same as $\boldsymbol{\Lambda}$. Simple Kriging weights $\boldsymbol{\Lambda}$ can be expressed as:

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_{DD}^{-1}\boldsymbol{\Sigma}_{DY} \Rightarrow \boldsymbol{\Lambda}^T = \boldsymbol{\Sigma}_{YD}\boldsymbol{\Sigma}_{DD}^{-1} \tag{5.2}$$

and according to Eq. 5.1 and 5.2, the SK target predictions can be expressed as:

$$\hat{\mathbf{y}} = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YD}\boldsymbol{\Sigma}_{DD}^{-1}[\mathbf{d} - \boldsymbol{\mu}_D]$$

In terms of target supports the SK target prediction $\hat{y}(t_m)$ at the m-th target support t_m is equal to:

$$\begin{aligned} \hat{y}(t_m) &= \mu_Y(t_m) + \sum_{n=1}^{N} \lambda_m(s_n)[d(s_n) - \mu_D(s_n)] \\ &= \mu_Y(t_m) + \lambda_m^T[d(s_n) - \mu_D(s_n)] = \mu_Y(t_m) + \boldsymbol{\sigma}_m^T\boldsymbol{\Sigma}_{DD}^{-1}[d(s_n) - \mu_D(s_n)] \end{aligned}$$

The $(M \times 1)$ unknown target attribute vector $\mathbf{y}$ contains M point support attribute values, and the $(N \times 1)$ source data vector $\mathbf{d}$ is assumed to be linked to these unknown point values via a $(N \times M)$ known sampling function array $\mathbf{G}$ as $\mathbf{d} = \mathbf{G}\mathbf{y}$:

$$\begin{bmatrix} d(s_1) \\ \vdots \\ d(s_n) \\ \vdots \\ d(s_N) \end{bmatrix} = \begin{bmatrix} g_1(t_1) & \cdots\cdots & g_1(t_m) & \cdots\cdots & g_1(t_M) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ g_n(t_1) & \cdots\cdots & g_n(t_m) & \cdots\cdots & g_n(t_M) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ g_N(t_1) & \cdots\cdots & g_N(t_m) & \cdots\cdots & g_N(t_M) \end{bmatrix} \begin{bmatrix} y(t_1) \\ \vdots \\ \vdots \\ y(t_m) \\ \vdots \\ \vdots \\ y(t_M) \end{bmatrix}$$

where the n-th row of $\mathbf{G}$ contains information on the contribution of the M unknown point values to the n-th known source datum: $d(s_n) = \sum_{m=1}^{M} g_n(t_m) y(t_m)$

For point support source data, $\mathbf{G}$ contains only 1 non-zero value in each of its N rows, meaning that N (out of M) target locations coincide with the N source data locations. For non-point support source data, there are multiple non-zero entries in each row of $\mathbf{G}$. The general statistics of source data and target values are shortly summarized as follows:

- The functional link between source data and target values is:

$$\mathbf{G}\mathbf{y} = \mathbf{d}$$

- Statistics of the N RVs pertaining to the source supports are linked to the statistics of the M RVs pertaining to the target supports
- The expectation vectors of the source support RVs is : $\boldsymbol{\mu}_D = \mathbf{G}\boldsymbol{\mu}_Y$
- Covariance matrices between target and source support RVs are: $\boldsymbol{\Sigma}_{YD} = \boldsymbol{\Sigma}_{YY}\mathbf{G}^T$
- Covariance matrices between source support RVs are: $\boldsymbol{\Sigma}_{DD} = \mathbf{G}\boldsymbol{\Sigma}_{YY}\mathbf{G}^T = \mathbf{G}\boldsymbol{\Sigma}_{YD}$.

Simple Kriging interpolation exactly reproduces the source data; the application of the definition of the source data $\mathbf{d} = \mathbf{G}\mathbf{y}$ to the vector $\hat{\mathbf{y}}$ of SK target predictions, yields back the original source data, i.e., $\mathbf{G}\hat{\mathbf{y}} = \mathbf{d}$ as follows:

$$\begin{aligned} \mathbf{G}\hat{\mathbf{y}} &= \mathbf{G}\left(\boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YD}\boldsymbol{\Sigma}_{DD}^{-1}[\mathbf{d} - \boldsymbol{\mu}_D]\right) = \mathbf{G}\boldsymbol{\mu}_Y + \mathbf{G}\boldsymbol{\Sigma}_{YD}\boldsymbol{\Sigma}_{DD}^{-1}[\mathbf{d} - \boldsymbol{\mu}_D] \\ &= \boldsymbol{\mu}_D + \boldsymbol{\Sigma}_{DD}\boldsymbol{\Sigma}_{DD}^{-1}[\mathbf{d} - \boldsymbol{\mu}_D] = \boldsymbol{\mu}_D + [\mathbf{d} - \boldsymbol{\mu}_D] = \boxed{\mathbf{d}} \end{aligned}$$

The data reproduction property of SK is independent of the covariogram model $\sigma_Y(\mathbf{h})$ used to populate $\boldsymbol{\Sigma}_{YY}$, hence $\boldsymbol{\Sigma}_{YD} = \boldsymbol{\Sigma}_{YY}\mathbf{G}^T$ and $\mathbf{G}\boldsymbol{\Sigma}_{YY}\mathbf{G}^T$; one need to only link consistently all covariance matrices to the point covariance matrix $\boldsymbol{\Sigma}_{YY}$. Kriging predictions reproduce (up to discretization approximations) any source datum that can be expressed as a linear combination of unknown point values; this includes point values, unequally weighted volume averages, as well as derivatives. Moreover, SK prediction errors at N

source supports are zeros; point source data are reproduced only if their sample locations coincide with prediction locations.

The $(M \times 1)$ vector $\mathbf{y}$ of true (unknown) target values can be decomposed in terms of the Kriging prediction $\hat{\mathbf{y}}$ (5.3 left) and the Kriging error (5.3 right) as:

$$\mathbf{y} = \hat{\mathbf{y}} + [\mathbf{y} - \hat{\mathbf{y}}] \tag{5.3}$$

where $[\mathbf{y} - \hat{\mathbf{y}}]$ is a $(M \times 1)$ vector of SK prediction errors. In general, SK prediction error $[\mathbf{y} - \hat{\mathbf{y}}]$ is not considered known, since $\mathbf{y}$ is unknown.

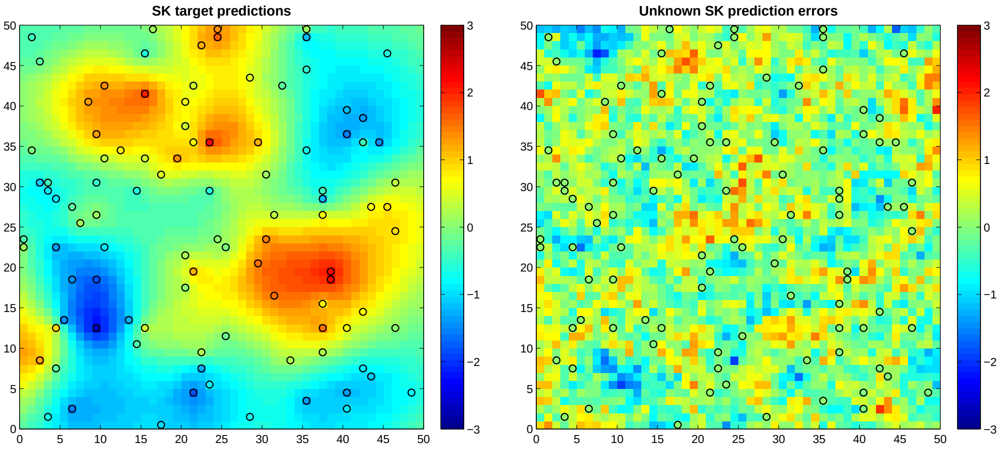


Figure 5.3: Kriging prediction realization (left) and Kriging error surface (right)

Only the N SK prediction errors at the N source supports are known and equal to 0, provided that the source data are error-free: $\mathbf{G}[\mathbf{y} - \hat{\mathbf{y}}] = [\mathbf{d} - \mathbf{d}] = \mathbf{0}$. However, the statistics (mean and covariance) of that error are considered known, since they are linked to the statistics of $\mathbf{y}$ which are assumed known. The expectation of SK predictor RVs is:

$$\begin{aligned} \boldsymbol{\mu}_{\hat{Y}} &= \mathbb{E}\{\hat{\mathbf{y}}\} = \mathbb{E}\{\boldsymbol{\mu}_Y\} + \boldsymbol{\Lambda}^T \mathbb{E}\{[\mathbf{d} - \boldsymbol{\mu}_D]\} \\ &= \boldsymbol{\mu}_Y + \boldsymbol{\Lambda}^T [\boldsymbol{\mu}_D - \boldsymbol{\mu}_D] = \boldsymbol{\mu}_Y \end{aligned}$$

and the expectation of SK prediction error RVs is:

$$\boldsymbol{\mu}_{Y-\hat{Y}} = \mathbb{E}\{\mathbf{y} - \hat{\mathbf{y}}\} = \mathbb{E}\{\mathbf{y}\} - \mathbb{E}\{\hat{\mathbf{y}}\} = \boldsymbol{\mu}_Y - \boldsymbol{\mu}_Y = \mathbf{0}$$

Prediction error RVs have zero expectations since Kriging is an unbiased predictor.

Conditional covariance $(M \times M)$ matrix $\boldsymbol{\Sigma}_{\hat{Y}\hat{Y}}$ of covariance values between all pairs of predictor RVs at all M target supports is equal to:

$$\begin{aligned} \boldsymbol{\Sigma}_{\hat{Y}\hat{Y}} &= \mathbb{E}\{[\hat{\mathbf{y}} - \boldsymbol{\mu}_{\hat{Y}}][\hat{\mathbf{y}} - \boldsymbol{\mu}_{\hat{Y}}]^T\} = \mathbb{E}\{\hat{\mathbf{y}}\hat{\mathbf{y}}^T\} - \boldsymbol{\mu}_Y \boldsymbol{\mu}_Y^T \\ &= \mathbb{E}\{(\boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YD}\boldsymbol{\Sigma}_{DD}^{-1}[\mathbf{d} - \boldsymbol{\mu}_D])(\boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YD}\boldsymbol{\Sigma}_{DD}^{-1}[\mathbf{d} - \boldsymbol{\mu}_D])^T\} - \boldsymbol{\mu}_Y \boldsymbol{\mu}_Y^T \\ &= \boldsymbol{\Sigma}_{YD}\boldsymbol{\Sigma}_{DD}^{-1}\mathbb{E}\{[\mathbf{d} - \boldsymbol{\mu}_D][\mathbf{d} - \boldsymbol{\mu}_D]^T\}\boldsymbol{\Sigma}_{DD}^{-1}\boldsymbol{\Sigma}_{DY} \\ &= \boldsymbol{\Sigma}_{YD}\boldsymbol{\Sigma}_{DD}^{-1}\boldsymbol{\Sigma}_{DY} \neq \boldsymbol{\Sigma}_{YY} \end{aligned}$$

whereas $(M \times M)$ matrix $\mathbf{\Sigma}_{Y\hat{Y}}$ of covariance values between all pairs of target and predictor RVs at all M target supports:

$$\begin{aligned}
\mathbf{\Sigma}_{Y\hat{Y}} &= \mathbb{E}\{[\mathbf{y}-\boldsymbol{\mu}_Y][\hat{\mathbf{y}}-\boldsymbol{\mu}_{\hat{Y}}]^T\} = \mathbb{E}\{\mathbf{y}\hat{\mathbf{y}}^T\} - \mathbb{E}\{\boldsymbol{\mu}_Y\boldsymbol{\mu}_{\hat{Y}}^T\} \\
&= \mathbb{E}\{\mathbf{y}(\boldsymbol{\mu}_Y + \mathbf{\Sigma}_{YD}\mathbf{\Sigma}_{DD}^{-1}[\mathbf{d}-\boldsymbol{\mu}_D])^T\} - \boldsymbol{\mu}_Y\boldsymbol{\mu}_Y^T \\
&= \mathbb{E}\{\mathbf{y}\}\boldsymbol{\mu}_Y^T + \mathbb{E}\{\mathbf{y}[\mathbf{d}-\boldsymbol{\mu}_D]^T\}\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{DY} - \boldsymbol{\mu}_Y\boldsymbol{\mu}_Y^T \\
&= \mathbf{\Sigma}_{YD}\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{DY} = \mathbf{\Sigma}_{\hat{Y}Y}
\end{aligned}$$

and $(M \times M)$ matrix $\mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})}$ of covariance values between all pairs of prediction error RVs at all M target supports:

$$\begin{aligned}
\mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})} &= \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{Y\hat{Y}} - \mathbf{\Sigma}_{\hat{Y}Y} + \mathbf{\Sigma}_{\hat{Y}\hat{Y}} \\
&= \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{Y\hat{Y}} = \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YD}\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{DY} \qquad (5.4)
\end{aligned}$$

entries on the diagonal of $\mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})}$ constitute $(M \times 1)$ vector of prediction error variances at all M target supports; m-th entry of that vector: $\sigma_Y(t_m) - \mathbf{w}_m^T\boldsymbol{\sigma}_m =$ Kriging error variance. Lastly, $(M \times M)$ matrix $\mathbf{\Sigma}_{\hat{Y}(Y-\hat{Y})}$ of covariance values between all pairs of predictor and prediction error RVs at all M target supports:

$$\mathbf{\Sigma}_{\hat{Y}(Y-\hat{Y})} = \mathbf{\Sigma}_{\hat{Y}Y} - \mathbf{\Sigma}_{\hat{Y}\hat{Y}} = \mathbf{0}$$

Predictor RVs and prediction error RVs are uncorrelated; homoscedasticity is a regression characteristic.

Simple Kriging predictions are optimal in the least squares sense as they are unbiased and have minimum prediction error variance among all other weighted linear combinations of data; Kriging is the Best Linear Unbiased Predictor (BLUP) and yields the set of most locally accurate predictions. Morevover the covariance matrix of prediction errors (as well as all other conditional covariance matrices) is homoscedastic, i.e., does not depend on the data values, but only on: (i) the configuration of source and target supports, and (ii) the covariance model $\sigma_Y(\mathbf{h})$. The true value $y(t_m)$ at target support t_m don't need to be known, in order to be able to calculate the associated prediction error variance $\hat{\sigma}(t_m)$; very important for sampling design purposes.

On the other hand, the covariance matrix $\mathbf{\Sigma}_{\hat{Y}\hat{Y}}$ of Kriging predictor RVs is not the same as the true covariance matrix $\mathbf{\Sigma}_{YY}$; Kriging predictions are not considered to provide structural accuracy, i.e., variogram reproduction. To complicate things, $\mathbf{\Sigma}_{\hat{Y}\hat{Y}}$ is non-stationary, i.e., the conditional covariance between two pairs of predictor RVs $\{\hat{Y}(t_m), \hat{Y}(t_{m'})\}$ and $\{\hat{Y}(t_{m''}), \hat{Y}(t_{m'''})\}$ at two pairs of target supports $\{t_m, t_{m'}\}$ and $\{t_{m''}, t_{m'''}\}$ separated by the same distance, is not the same. In spectral analysis terminology, Kriging is a non-stationary low-pass filter. Furthermore, the conditional covariance matrix $\mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})}$ of prediction error RVs is indeed homoscedastic only in the multivariate Gaussian case, i.e., when the joint distribution of all N source RVs and all M target RVs is a $(N+M)$-variate Gaussian distribution.

Fortunately Kriging is used as an intermediate step in stochastic simulation, to condition Gaussian realizations to known measurements (conditional simulation); this can be expressed as: $\mathbf{y} \sim N(\boldsymbol{\mu}_{Y|\mathbf{d}}, \mathbf{\Sigma}_{YY|\mathbf{d}})$ instead of $\mathbf{y} \sim N(\boldsymbol{\mu}_Y, \mathbf{\Sigma}_{YY})$ where $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ denotes the multivariate Gaussian (normal) distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$.

5.3 Conditional Latin Hypercube simulation: Algorithm 1

According to the previous subsection, parameter $[\mathbf{y} - \hat{\mathbf{y}}]$ of equation 5.3, can now be re-expressed as a $(M \times 1)$ vector of zero-mean SK prediction errors. Moreover, since $\hat{\mathbf{y}}$ is independent from $\mathbf{y} - \hat{\mathbf{y}}$, the target covariance matrix can be decomposed as:

$$\mathbf{\Sigma}_{YY} = \mathbf{\Sigma}_{\hat{Y}\hat{Y}} + \mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})}$$

Conditional attribute realizations in the multivariate Gaussian case can be constructed by: (i) performing Kriging to obtain the SK predictions $\hat{\mathbf{y}}$, (ii) generating a realization $[\mathbf{y} - \hat{\mathbf{y}}]$ from the SK prediction error random vector, and (iii) adding that realization to the SK predictions $\hat{\mathbf{y}}$ to obtain a conditional realization $\mathbf{y}_c$. The issue is how to generate a realization $[\mathbf{y} - \hat{\mathbf{y}}]$ from the multivariate Gaussian distribution of the SK prediction error random vector. One of the oldest and commonly used conditional simulation algorithms is the 'Direct Simulation of Kriging Prediction Error' (Journel, 1974). The relevant flowchart can be described as:

1. Compute Kriging predictions $\hat{\mathbf{y}}$; Fig. 5.4, left.

2. Compute conditional covariance matrix of Kriging prediction error RVs $\mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})}$; Fig. 5.4, right

3. Simulate, independently of the Kriging component $\hat{\mathbf{y}}$, a zero-mean realization $[\mathbf{y}-\hat{\mathbf{y}}]$ of the Kriging error component via Cholesky decomposition of the Kriging error covariance matrix $\mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})}$. Since SK is an exact interpolator, the simulated SK prediction errors at all N source supports are zeros: $\mathbf{G}[\mathbf{y} - \hat{\mathbf{y}}] = \mathbf{0}$; Fig. 5.5, right. Per the theory of Cholesky simulation, the simulated error realizations have covariance $\mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})}$.

4. Obtain a conditional realization as: $\mathbf{y}_q = \hat{\mathbf{y}} + [\mathbf{y} - \hat{\mathbf{y}}]$; Fig. 5.5, left.

5. Repeat steps 2 and 3 S times, to generate S simulated vectors $\{\mathbf{y}_q^{(s)}, s = 1, \ldots, S\}$ of target values with mean vector $\simeq \boldsymbol{\mu}_{Y|\mathbf{d}}$ and covariance matrix $\simeq \mathbf{\Sigma}_{YY|\mathbf{d}}$

6. Transform this conditional Simple Random sample into a conditional Latin Hypercube sample of size S using equation:

$$[y_q^{(s)}(\mathbf{c}_m)]_{LS} = G^{-1}\left(\frac{\text{rank}(y_q^{(s)}(\mathbf{c}_m)) - v_q^{(s)}(\mathbf{c}_m)}{S}; \hat{\mu}(\mathbf{c}_m), \hat{\sigma}^2(\mathbf{c}_m)\right), \quad s = 1, \ldots, S \tag{5.5}$$

where G^{-1} is the inverse CDF of the local Gaussian conditional distribution of RV $Y_q(\mathbf{c}_m)$ with parameters the local conditional estimation provided by the Kriging prediction $\hat{\mu}(\mathbf{c}_m)$ and the local conditional variance provided by the Kriging prediction error variance $\hat{\sigma}^2(\mathbf{c}_m)$. Term $\text{rank}(y_q^{(s)}(\mathbf{c}_m))$ denotes the rank of the s-th value $y_q^{(s)}(\mathbf{c}_m)$ in the conditional SR sample $\mathbf{y}_q(\mathbf{c}_m)$ at location $\mathbf{c}_m$.

In essence, Eq. (5.5) involves a set of S spatially correlated probability values (the argument of the inverse Gaussian CDF), which are also marginally (location-wise) stratified. The rank value $\text{rank}(y_q^{(s)}(\mathbf{c}_m))$ identifies the probability stratum associated with an original value $y_q^{(s)}(\mathbf{c}_m)$ of the SR sample at location $\mathbf{c}_m$. The addition of a random number $v_q^{(s)}(\mathbf{c}_m)$ uniformly distributed in $[0, 1]$ furnishes a random probability perturbation within than stratum. Those stratified probability values are then used to derive the corresponding stratified Gaussian quantiles via the inverse local CDF $G^{-1}\left(y_q; \hat{\mu}(\mathbf{c}_m), \hat{\sigma}^2(\mathbf{c}_m)\right)$. Note that spatial correlation is induced in the Latin hypercube realizations via the ranks of the original (generated via SR sampling) conditional simulations.

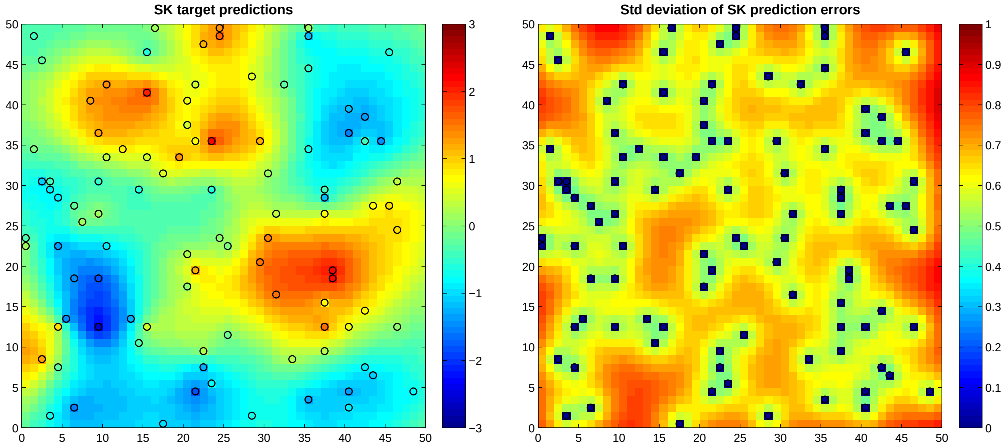


Figure 5.4: Simple Kriging predictions (left) and error std deviations (right)

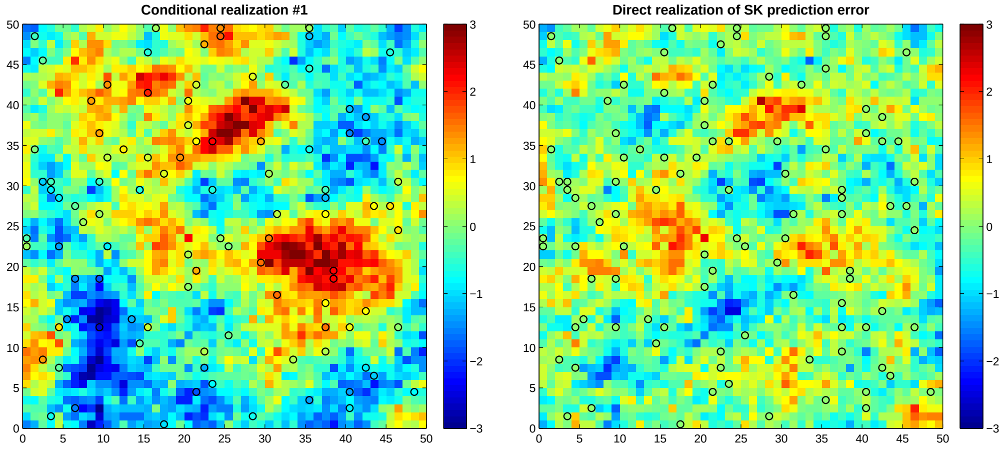


Figure 5.5: Conditional simulation (left) and prediction error (right)

In a nutshell the above steps 1 to 5 of the flowchart can be expressed according to Alabert (1987) and Davis (1987) as the reproduction of a conditional simulation $\mathbf{y}_q$ on the M target supports in a Gaussian or transformed Gaussian setting can be expressed as:

$$\mathbf{y}_q = \mathbf{\Sigma}_{YD}\mathbf{\Sigma}_{DD}^{-1}\mathbf{d} + \hat{\mathbf{L}}\mathbf{w} = \text{(simple Kringing value)} + \text{(simulated Kriging error)} \quad (5.6)$$

where $\mathbf{\Sigma}_{YD}\mathbf{\Sigma}_{DD}^{-1} = \mathbf{\Lambda}$, and $\mathbf{\Lambda}$ denotes the Kriging weights of the system, $\hat{\mathbf{L}}$ is the Cholesky decomposition factor of $\mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})}$ and according to 5.4, $\mathbf{\Sigma}_{(Y-\hat{Y})(Y-\hat{Y})} = \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YD}\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{DY}$, and $\mathbf{w}$ is $(M \times 1)$ vector $\mathbf{w} = [w(t_m), m = 1, \ldots, M]^T$ of random values at M target supports. $\mathbf{\Sigma}_{YD}\mathbf{\Sigma}_{DD}^{-1}\mathbf{d}$ is the component accounting for the conditioning data and $\hat{\mathbf{L}}\mathbf{w}$ the random component that serves for generating multiple realizations. For a more detailed and comprehensive exposition of the conditional simulation via direct simulation of the Kriging error, the reader is referred to (Alabert, 1987).

In the multivariate Gaussian case, the ensemble mean and variance of target realizations converge to Kriging predictions and variances, as the number of realizations increases. The shortcoming is that since the algorithm is based on the Cholesky decomposition, it works for not too many target supports, $M < 10,000$ or so.

5.4 Condtional Latin Hypercube simulation: Algorithm 2

A more widely applicable approach in terms of possible number of target supports $M >> 10,000$ is the conditional simulation via the indirect simulation of Kriging prediction error (Chilès and Delfiner (1999); Goovaerts (1997)); true value decomposition of Eq. 5.3. The basic difference is the simulation of a realization of the Kriging error by simulating an unconditional realization and using simulated source data extracted from this realization for the simulation of the Kriging error; simulated source data are extracted with the same sampling mechanism (same supports) as the original source data.

The basic concept is the simulation of a realization of the Kriging error $[\mathbf{y} - \hat{\mathbf{y}}]$ by: (i) simulating a realization $\mathbf{y}$; this realization is unconditional and has covariance $\mathbf{\Sigma}_{YY}$, e.g., it is generated using a stationary covariogram model $\sigma_Y(\mathbf{h})$, and (ii) repeating the Kriging procedure using simulated source data extracted from the realization $\mathbf{y}$. Simulated Kriging should be performed using the same covariogram model used to generate the unconditional realization $\mathbf{y}$, as well as to compute the original Kriging predictions $\hat{\mathbf{y}}$; simulated source data are extracted with the same sampling mechanism (same supports) as the original source data.

Thus, a conditional realization i.e., a $(M \times 1)$ vector $\mathbf{y}_q$ of conditionally simulated target values can be obtained as:

$$\mathbf{y}_q = \hat{\mathbf{y}} + [\mathbf{y} - \hat{\mathbf{y}}_{sim}] \quad (5.7)$$

where $\hat{\mathbf{y}}$ is a $(M \times 1)$ vector of Kriging predictions, $\mathbf{y}$ is a $(M \times 1)$ vector of unconditionally simulated target values, and $\hat{\mathbf{y}}_{sim}$ is a $(M \times 1)$ vector of simulated Kriging predictions, i.e., Kriging predictions with simulated source data. The algorithm flowchart can be described as:

1. Generate a $(M \times 1)$ vector $\mathbf{y}$ of simulated target values with covariance matrix $\mathbf{\Sigma}_{YY}$ using unconditional simulation, Fig. 5.6 (left); there exist efficient algorithms (e.g. FFT-based method presented in subsection 4.1.3) to do so especially for very large simulation domains.

2. Generate the $(M \times 1)$ vector $\mathbf{d}$ of simulated source data as: $\mathbf{d} = \mathbf{G}\mathbf{y}$, Fig. 5.6 (right); per stationarity and ergodicity, the mean of the simulated data is approximately $\boldsymbol{\mu}_Y$, and their $(N \times N)$ covariance matrix is approximately $\mathbf{\Sigma}_{DD}$.

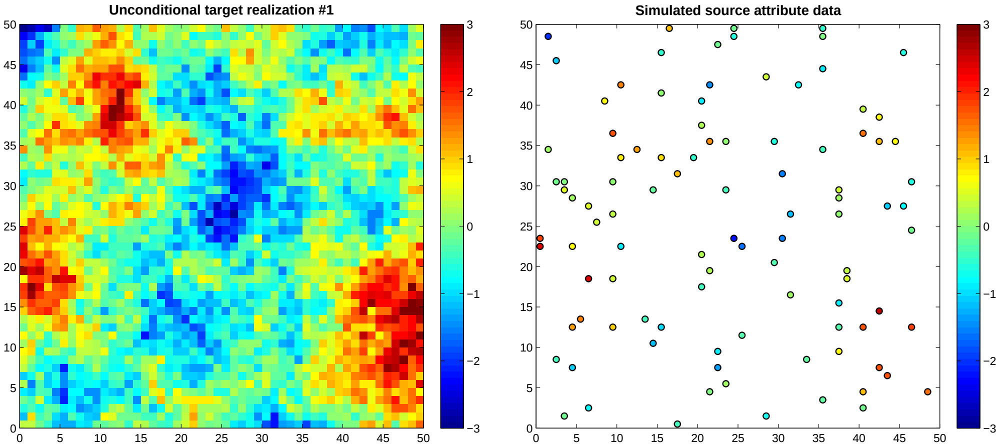


Figure 5.6: Unconditional target realization (left) and simulated source data (right)

3. Perform Simple Kriging using the simulated data vector $\mathbf{d}$ to obtain a $(M \times 1)$ vector $\hat{\mathbf{y}}_{sim}$ of simulated SK predictions (Fig. 5.7, left) as:

$$\hat{\mathbf{y}}_{sim} = \boldsymbol{\mu}_Y + \mathbf{W}^T[\mathbf{d} - \boldsymbol{\mu}_D] = \boldsymbol{\mu}_Y \mathbf{\Sigma}_{YD} \mathbf{\Sigma}_{DD}^{-1}[\mathbf{d} - \boldsymbol{\mu}_D]$$

   Since the source support configuration remains the same, and all conditional covariance matrices are homoscedastic, the simulated Kriging realization has the same properties as the original Kriging; simulated Kriging predictions reproduce the simulated source data: $\mathbf{d} = \mathbf{G}\hat{\mathbf{y}}$.

4. Compute simulated Kriging prediction errors: $[\mathbf{y} - \hat{\mathbf{y}}_{sim}]$, Fig. 5.7 (right). Per the data reproduction property of Kriging, the simulated Kriging prediction error is zero at any source support: $\mathbf{d} - \mathbf{G}\hat{\mathbf{y}} = \mathbf{0}$

5. Obtain conditional realization as: $\mathbf{y}_q = \hat{\mathbf{y}} + [\mathbf{y} - \hat{\mathbf{y}}_{sim}]$. Since the SK predictions reproduce the source data at the source supports, the simulated SK prediction errors are zero at those supports, hence the original source data are reproduced, i.e.: $\mathbf{d} = \mathbf{G}\mathbf{y}_q$

6. Transform this conditional Simple Random sample into a conditional Latin Hypercube sample of size S using Eq. 5.5.

A representation of the above flowchart, showing the steps for generating a conditional realization via indirect simulation of the Kriging error is depicted in Fig. 5.8.

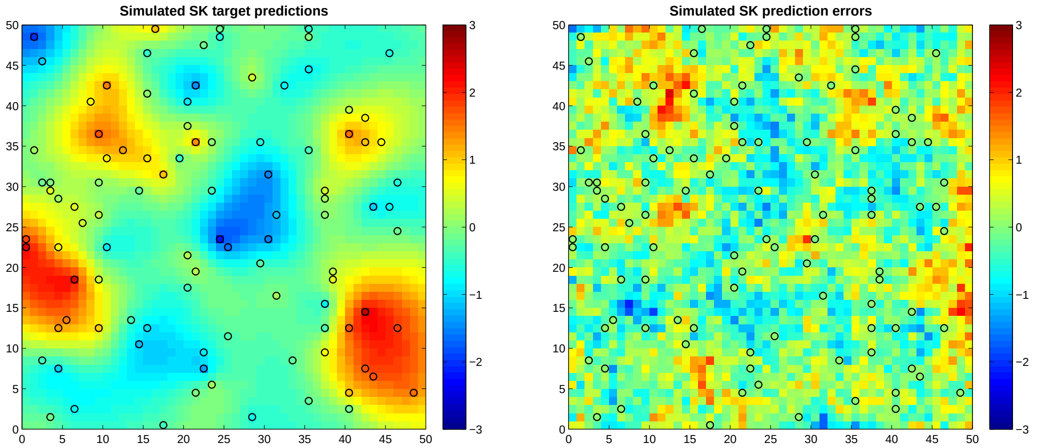


Figure 5.7: Simulated SK predictions (left) and simulated SK prediction errors (right)

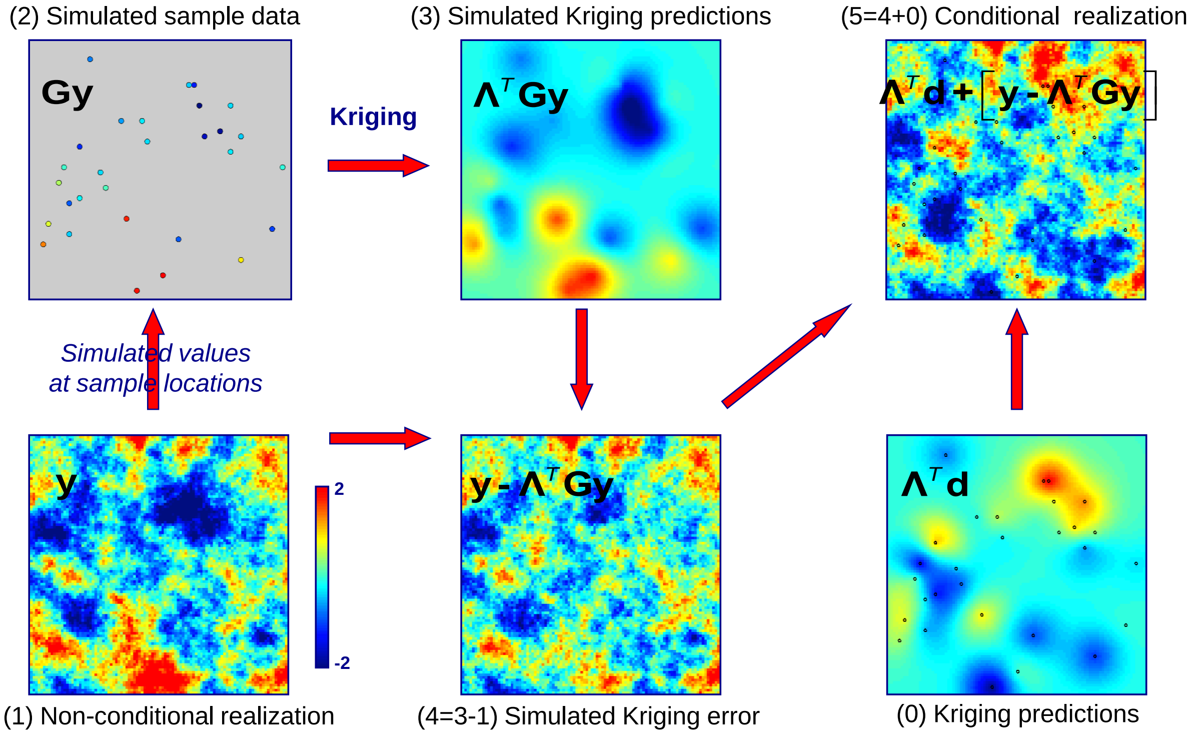

(2) Simulated sample data
Gy
Kriging
(3) Simulated Kriging predictions
$\Lambda^T Gy$
(5=4+0) Conditional realization
$\Lambda^T d + [y - \Lambda^T Gy]$
Simulated values
at sample locations
y
2
-2
$y - \Lambda^T Gy$
$\Lambda^T d$
(1) Non-conditional realization
(4=3-1) Simulated Kriging error
(0) Kriging predictions

Figure 5.8: Conditional simulation via indirect simulation of the Kriging error flowchart

Recap

Kriging provides the best linear prediction at any single target support and the conditional covariance matrix of the prediction error RVs, whose diagonal elements are the classical Kriging error variances. Kriging also reproduces any source datum that can be expressed as a linear combination of unknown point values, as long as one consistently links all required covariances to a point covariance matrix. Stochastic conditional simulation can be regarded as the task of constructing realizations of the (zero-mean) Kriging prediction error random vector and adding them to the Kriging predictions; the statistics of this random vector can be computed from those of the original attribute random vector.

There are two ways to construct realizations of the Kriging prediction error random vector: (i) direct simulation of the Kriging prediction error via Cholesky decomposition of the conditional covariance matrix; prohibitive for very large simulation grids, and (ii) indirect simulation of the Kriging prediction error by performing Kriging with simulated data constructed from unconditionally simulated attribute realizations; the latter, further employed in next section 5.5 for combining conditional sampling with LH sampling from random fields on very large grids, allows for more efficient computations when simulation locations coincide with the nodes of a regular grid of possibly million nodes.

5.5 Conditional Latin hypercube simulation of random fields on large grids

Applying conditional Latin Hypercube simulation of random fields on very large grids is actually a combination of: (i) simulation on very large regular grids, Sec. 4.1.3, (ii) Latin Hypercube sampling from random fields, Sec. 2.3.2, and (iii) conditional simulation via the indirect simulation of Kriging error; algorithm 2, Sec. 5.4.

The reference order of the above three algorithms is not random, since one has to first simulate unconditional Simple Random realizations on very large grids, then apply the post-processing transformation of these conditional Simple Random realizations to Latin Hypercube realizations by marginally stratifying its entries and finally condition these realizations on the available data (wells in the three dimensional case study adopted). More specifically, the algorithm flowchart can be described as:

1. Equation 4.5 (Dietrich, 1997) is firstly employed S times, for generating a $(\tilde{S} \times M)$ matrix $\tilde{\mathbf{Y}}$ with simulated (possibly transformed) Gaussian deviates at the nodes of the extended grid, from which one extracts the $(S \times M)$ matrix $\mathbf{Y}$ of simulated values at the original M grid nodes.

2. Following, Eq. 2.14 is employed to transform the previously generated S conditional simple random realizations on very large grids, to S Latin hypercube realizations, from which one extracts the $(S \times M)$ matrix $\mathbf{Y}$ of simulated values at the original M grid nodes.

3. Last, Eq. 5.7 is adopted to further condition the $\mathbf{Y}$ unconditional LH realizations on the $(N \times S)$ conditioning data.

The above algorithm sequence can be easily comprehended through the following Figures holding the pictorial representation of the above three steps.

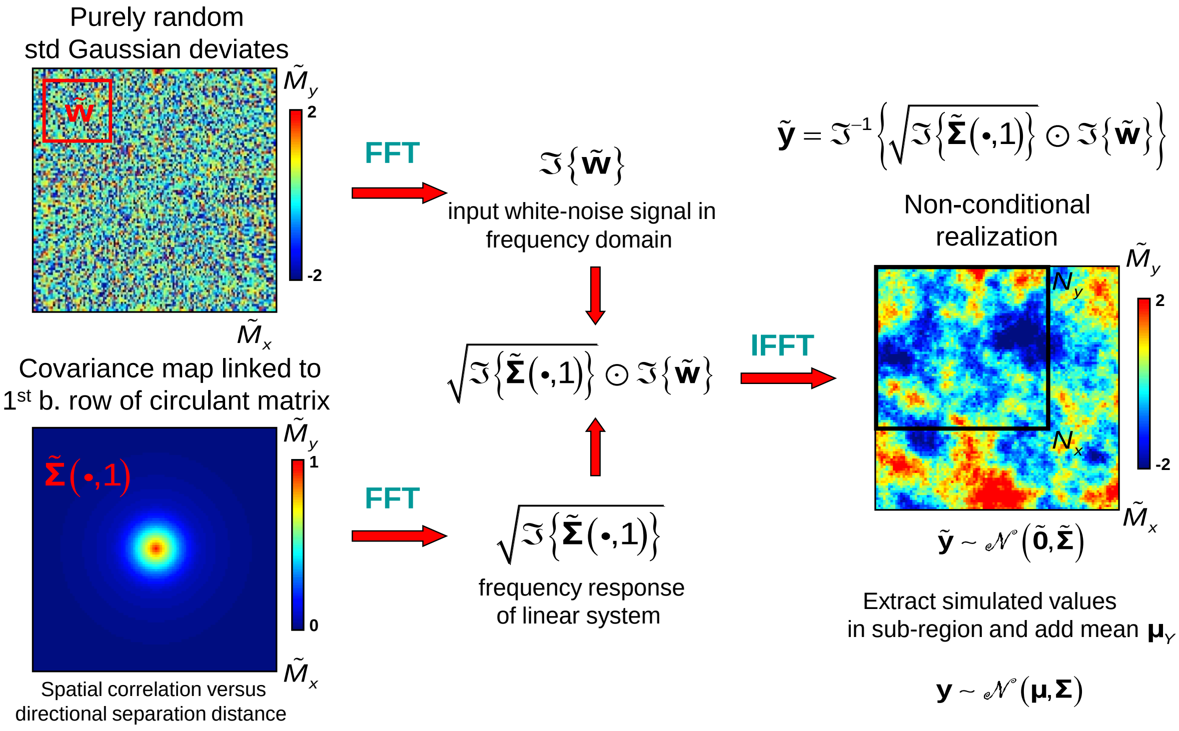

Purely random
std Gaussian deviates
$\tilde{\mathbf{w}}$
$\tilde{M}_y$
2
-2
$\tilde{M}_x$
FFT
$\Im\{\tilde{\mathbf{w}}\}$
input white-noise signal in
frequency domain
$\tilde{\mathbf{y}} = \Im^{-1}\left\{\sqrt{\Im\{\tilde{\boldsymbol{\Sigma}}(\bullet,1)\}} \odot \Im\{\tilde{\mathbf{w}}\}\right\}$
Non-conditional
realization
$\tilde{M}_y$
N_y
2
-2
N_x
$\tilde{M}_x$
$\sqrt{\Im\{\tilde{\boldsymbol{\Sigma}}(\bullet,1)\}} \odot \Im\{\tilde{\mathbf{w}}\}$
IFFT
Covariance map linked to
1st b. row of circulant matrix
$\tilde{\boldsymbol{\Sigma}}(\bullet,1)$
$\tilde{M}_y$
1
0
$\tilde{M}_x$
FFT
$\sqrt{\Im\{\tilde{\boldsymbol{\Sigma}}(\bullet,1)\}}$
frequency response
of linear system
$\tilde{\mathbf{y}} \sim \mathcal{N}(\tilde{\mathbf{0}},\tilde{\boldsymbol{\Sigma}})$
Extract simulated values
in sub-region and add mean $\boldsymbol{\mu}_Y$
$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$
Spatial correlation versus
directional separation distance

Figure 5.9: Schematic example illustrating the generation of one realization from a Gaussian random field model on a large grid; see text for details.

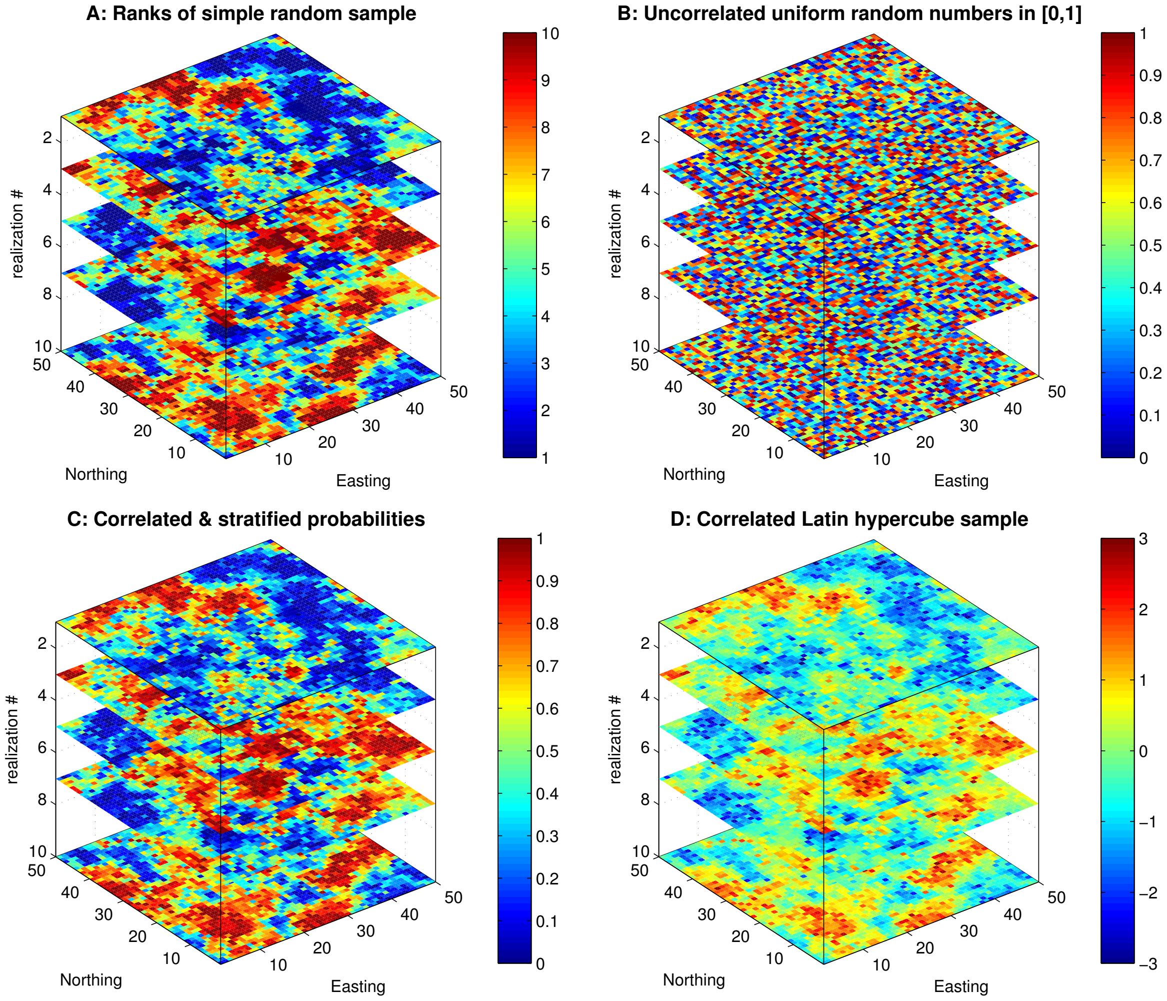


Figure 5.10: **A:** Rank ordered version of SR sample of Fig. 2.14, **B:** Uniform random numbers in $[0, 1]$, simulated independently at each grid node, **C:** Correlated probability values, stratified at each grid node, derived from **A** and **B**, **D:** Final LH sample of size $S = 10$, whose values are derived as quantiles of a standard Gaussian RV for the stratified probabilities in **C**.

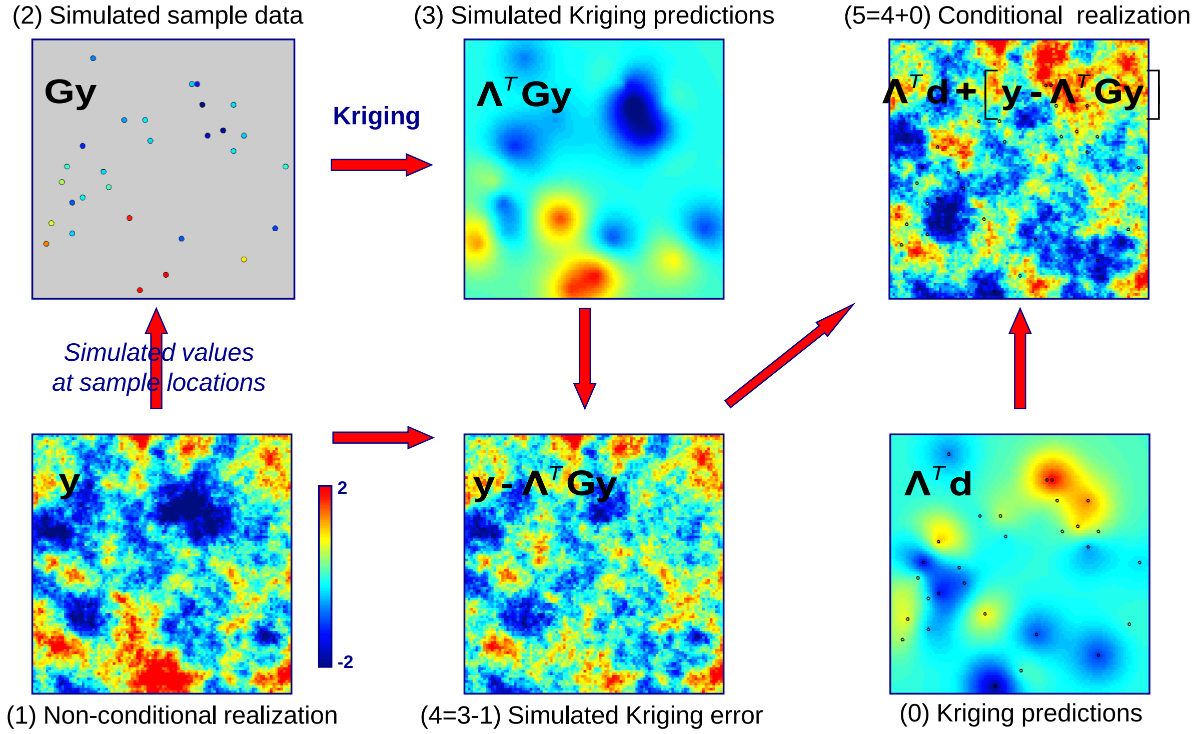

(2) Simulated sample data
Gy
Kriging
(3) Simulated Kriging predictions
$\Lambda^T Gy$
(5=4+0) Conditional realization
$\Lambda^T d + [y - \Lambda^T Gy]$
Simulated values
at sample locations
y
2
-2
$y - \Lambda^T Gy$
$\Lambda^T d$
(1) Non-conditional realization
(4=3-1) Simulated Kriging error
(0) Kriging predictions

Figure 5.11: Conditional simulation via indirect simulation of the Kriging error flowchart

5.6 3D case study

This Section presents a three dimensional case study on a very large (> 1000000 nodes) grids, comparing the proposed conditional LH sampling to SR sampling from a lognormal random field. Both conditional sampling methods model the spatial distribution of saturated hydraulic conductivity in hydrogeological context involving flow and transport in a partial heterogeneous porous medium by employing indirect sampling of the Kriging error (Sec. 5.4). Direct sampling of the Kriging error (Sec. 5.3) cannot be adopted for this case study since it uses Cholesky decomposition which is prohibitive for large grids. The comparison between the two methods is performed along the lines of Sec. 2.4. More precisely, the two sampling methods are firstly evaluated in terms of their ability to reproduce ensemble statistics of the random field; these statistics are computed from a very large set of conductivity realizations generated via simple random sampling. That large set of conductivity fields is used to derive a large set of solute concentration fields through flow and transport simulation; the two sampling methods are then compared again in terms of their ability to reproduce ensemble statistics of the solute concentration field. Fig. 5.12 illustrates a pictorial representation of a three dimensional simulated conductivity field (left) and the respective concentration realization (right).

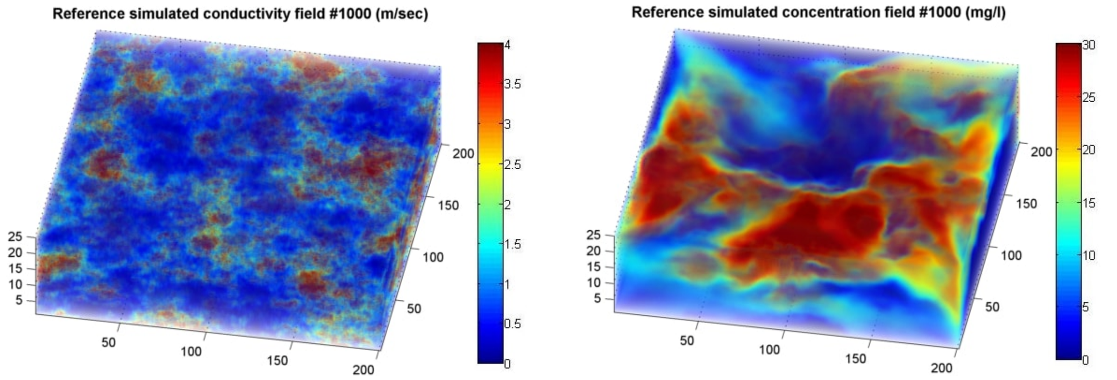

Reference simulated conductivity field #1000 (m/sec)
4
3.5
3
2.5
2
1.5
1
0.5
0
25
20
15
10
5
50
100
150
200
50
100
150
200
Reference simulated concentration field #1000 (mg/l)
30
25
20
15
10
5
0
25
20
15
10
5
50
100
150
200
50
100
150
200

Figure 5.12: A realization of a 3D lognormal random field (left), and corresponding 3D solute concentration realization (right).

The following two subsections present a synthetic case study involving a three dimensional groundwater flow system, discretized using a very large grid (> 1.000.000 nodes). Simple random (SR) and Latin hypercube (LH) sampling are considered for generating realizations from a lognormal random field, modeling the spatial distribution of saturated hydraulic conductivity in a hydrogeological flow and transport problem.

5.6.1 Hydraulic conductivity

The dimensions of the flow system are 1005 m by 1005 m by 125 m, discretized into a 201 × 201 × 25 grid with uniform cells of size 5 m by 5 m. Porosity is assumed constant throughout the domain and equal to 0.25. No flow conditions ($\partial h/\partial n = 0$) are assigned to the rest of the domain boundaries, and the Modflow software McDonald and Harbaugh (1988) is used to obtain the steady state flow solution.

Hydraulic coductivity reference ensemble statistics are derived from a set of 1000 realizations generated via SR sampling. The reference statistics for hydraulic conductivity fields consist of: (i) the ensemble average field shown in Fig. 5.13 (left), and (ii) the ensemble standard deviation field shown in Fig. 5.13 (right).

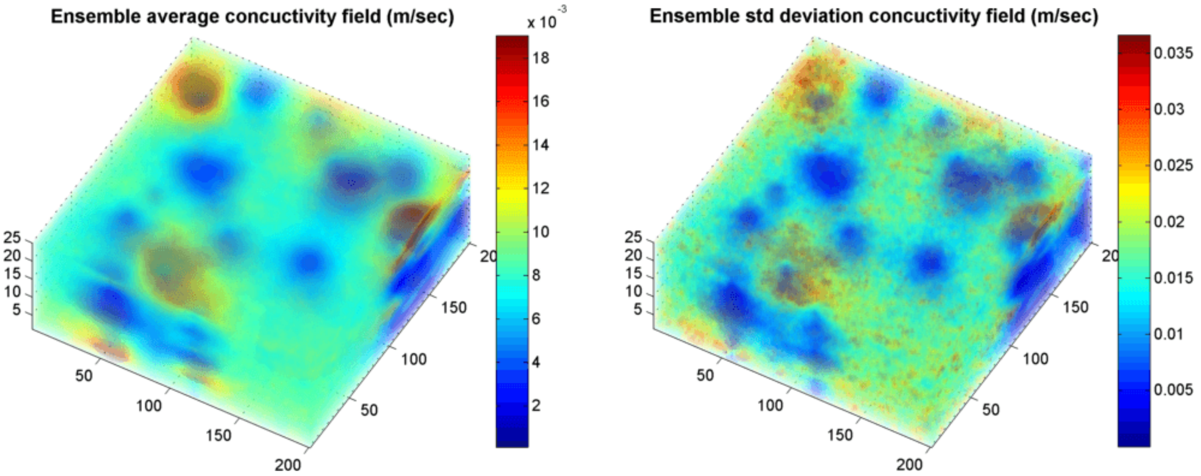


Figure 5.13: Reference ensemble average (left) and standard deviation (right) conductivity fields, computed from 1000 hydraulic conductivity realizations generated via simple random sampling.

In terms of sample size or number of realizations per method, three such sizes are considered; namely, $S = 20, 50$, and 80. Once a sample, say of size $S = 20$, is generated, the discrepancy between the statistics of the simulated ensemble and the reference statistics is quantified using the root mean squared error (RMSE). The computation of such error statistic is repeated over a set of $I = 100$ batches of realizations, with each batch containing the same sample size, $S = 20$ for example, thus estimating the sampling distributions of RMSE values for each sample size and for each method; these distributions are presented in terms of their means and medians, as well as their 75% and 95% probability intervals.

Fig. 5.14 depicts the reproduction of the ensemble average and standard deviation of hydraulic conductivity fields for the two sampling methods and the three sample sizes considered. It is easily appreciated from both figures that LH yields a better reproduction of ensemble average and standard deviation fields than SR sampling.

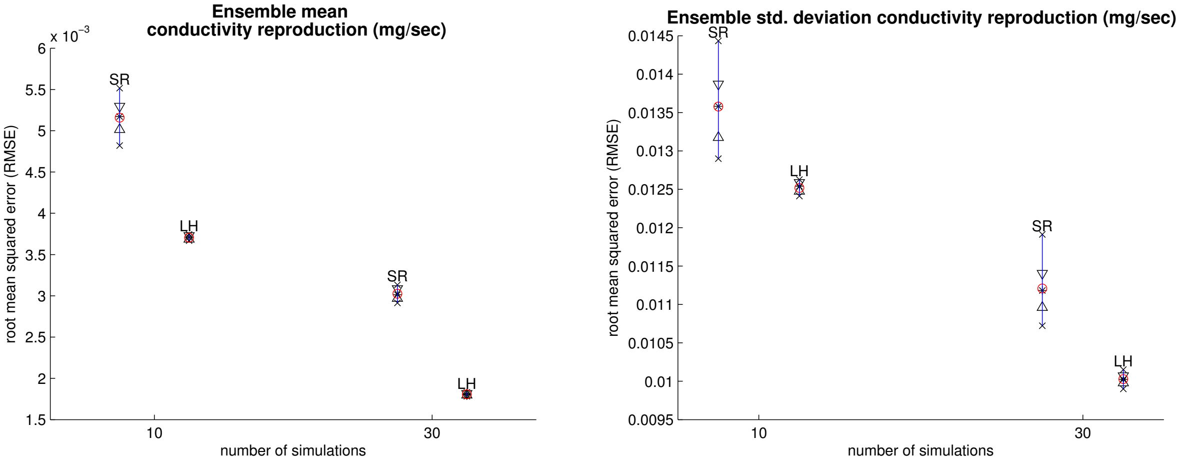


Figure 5.14: Reproduction of reference ensemble mean (left) and standard deviation (right) field for hydraulic conductivity; see text for details.

5.6.2 Solute concentration

For the solute transport problem, an initial concentration equal to 0 mg/l is assumed throughout the model domain. At time $t = 0$, a contaminant is introduced at the central cell (middle of three dimensional domain), along the upstream constant head boundary, with constant concentration $C_0 = 100$ mg/l. MT3D software Zheng (1990) is further used to solve the transport problem.

Here again, reference ensemble statistics are derived from the set of 1000 realizations of solute concentration fields corresponding to the 1000 realizations of hydraulic conductivity generated via SR sampling in the previous section. Reference statistics for solute concentration consist of: (i) the ensemble average field shown in Fig. 5.15 (left), and (ii) the ensemble standard deviation field shown in Fig. 5.15 (right).

Fig. 5.16 depicts the reproduction of the ensemble average and standard deviation concentration fields for the two sampling methods and the three sample sizes considered. LH sampling here again yields a better reproduction of these ensemble average and standard deviation fields than SR sampling.

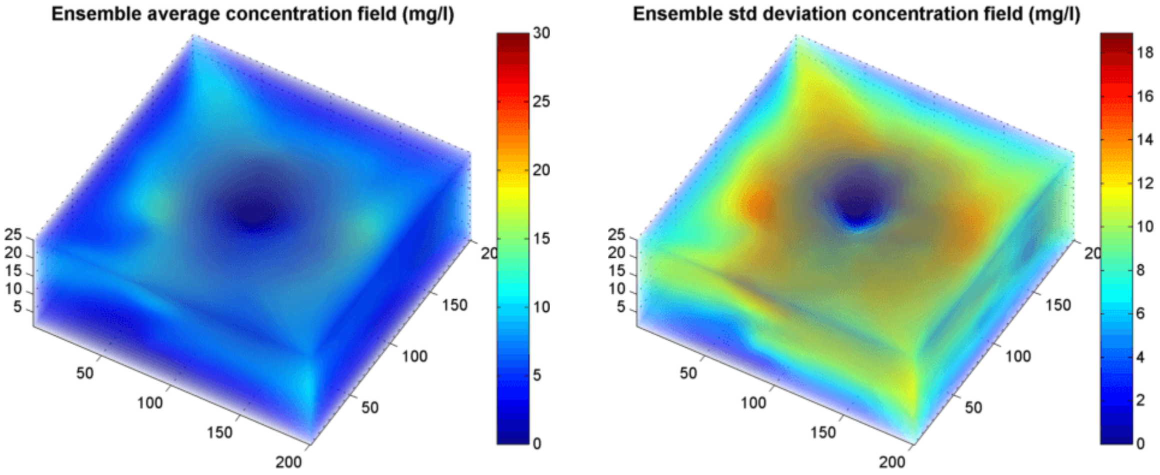

Ensemble average concentration field (mg/l)
Ensemble std deviation concentration field (mg/l)
30
25
20
15
10
5
0
18
16
14
12
10
8
6
4
2
0
25
20
15
10
5
50
100
150
200
20
150
100
50

Figure 5.15: Reference ensemble average (left) and standard deviation (right) concentration fields, computed from 1000 concentration realizations derived from 1000 hydraulic conductivity realizations generated via simple random sampling.

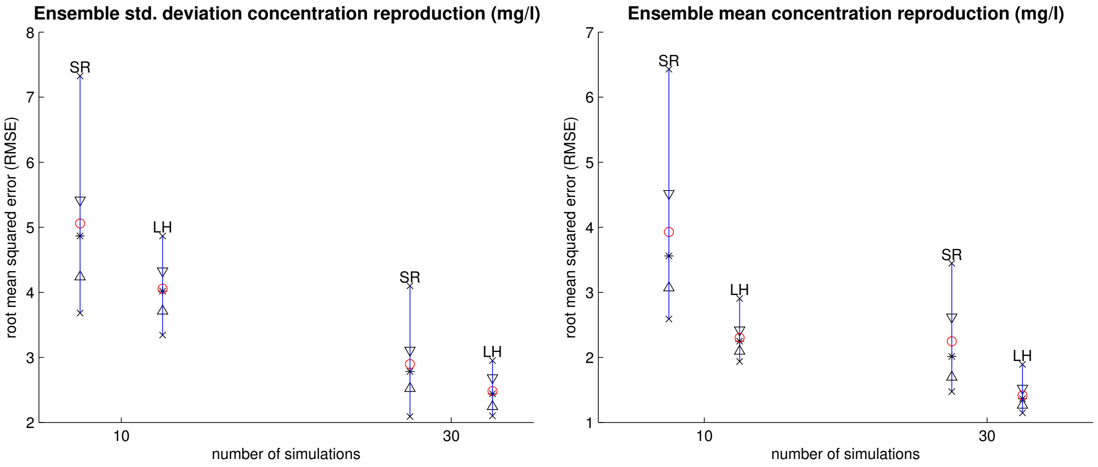


Figure 5.16: Reproduction of reference ensemble average (left) and standard deviation (right) field for concentration; see text for details.

5.7 Conclusions and Discussion

This chapter proposes a novel conditional simulation method for generating Latin hypercube samples from second order stationary Gaussian random fields at very large grids (with millions of nodes). LH sampling have been previously applied in an unconditional simulation context, and specifically for improved uncertainty analysis in hydrogeological settings. To the authors' knowledge, however, and apart from very few exceptions involving sequential Gaussian simulation, there is no published work investigating the application of LH sampling in a conditional geostatistical simulation setting, and in particular how LH relates to uncertainty analyses in hydrogeological investigations. The contribution of this chapter is precisely to propose methodological modifications so that LH sampling be adapted for geostatistical conditional simulation on very large application domains. To that respect, one could also suggest that this chapter proposes a novel algorithm for efficient conditional simulation of Gaussian random fields. More specifically, the proposed conditional LH sampling method combines an existing simulation method for creating SR samples from second order stationary random fields in large regular grids Dietrich (1997), with the LH sampling method of (Stein, 1987).

This work adapted LH sampling in a conditional simulation context. LH was contrasted with SR in terms of its ability to generate representative conditional realizations from a random field model, with term representative implying realizations that span efficiently the range of possible attribute outcomes corresponding to a conditional multivariate, (log)normal in this case, probability distribution. The performance of LH was evaluated in a hydrogeological context via a simple synthetic case study involving flow and transport in a heterogeneous porous medium. More precisely, the reproduction of reference ensemble statistics of a lognormal hydraulic conductivity field from conditional conductivity realizations generated via LH, was evaluated in the first part of the case study. The statistics considered included the ensemble mean and standard deviation. The reproduction of the same statistics for the ensemble concentration field resulting from solving a flow and transport boundary value problem for each hydraulic conductivity realization, was also evaluated in the second part of the case study.

As expected for the case of hydraulic conductivity, LH sampling achieved the best reproduction of marginal reference statistics for the ensemble mean and standard deviation followed by SR sampling (Fig. 5.14). In terms of solute concentration, LH sampling methods showed significantly better reproductions of the ensemble mean and standard deviation concentration fields than SR sampling (Fig. 5.16).

Chapter 6

Conclusions and future work

Uncertainty analysis of the outputs of each and every model is a critical requirement in each model's impact assessment and risk-consious policy decision-making. Uncertainty analysis in models with spatially distributed parameters specifically, is often performed within a Monte Carlo context in various engineering and scientific disciplines, including earth and environmental sciences (Caers, 2011). Along these lines, the propagation of the inherent spatial variability of petrophysical properties, such as hydraulic conductivity, to the predictions of physically-based, numerical models simulating flow and transport is an all-important task in hydrogeological investigations. In a classical, two-point, geostatistical context, spatial attribute variability is typically characterized via a random field model, log-normal in the case of hydraulic conductivity, which is often parameterized by a spatial constant mean, and a permissible variogram model. Simulations from such a random field, possibly reproducing known conductivity values at sample locations, are typically employed, along with the numerical simulators of flow and transport, in a Monte Carlo framework for evaluating, for example, the uncertainty in the spatial distribution of solute concentration. The objective is to cover the range of possible outcomes of the model at hand, and use that range of model outcomes to inform pertinent impact assessment studies and risk-conscious decision-making. Geostatistical simulation is typically performed using simple random (SR) sampling, a procedure that calls for the repeated evaluation of the physical model on a large number of input parameter realizations to arrive at a representative distribution of model outputs.

Latin hypercube sampling. The procedure of constructing a representative distribution of model outputs, however, can quickly become expensive in terms of both time and computer resources, particularly for the case of complex environmental models (Helton and Davis, 2002). More efficient alternatives to classical Monte Carlo simulation based on Simple Random (SR) sampling are stratified sampling methods, including Latin hypercube (LH) sampling, aiming at generating representative samples or realizations from a set of random variables with a given multivariate probability distribution. Here, term representative implies realizations spanning efficiently the range of possible attribute realizations corresponding to that probability distribution. In a more comprehensive Monte Carlo framework for uncertainty analysis, however, one should account for the functional links between model parameters and model outputs in the quest for representative samples or realizations (Caers, 2011). Remaining within the class of sampling approaches that do not utilize the information provided by the model itself, LH sampling has been shown to lead to model outputs with smaller sampling variability in their statistics than SR sampling for the same number of input simulated realizations; that efficiency, however, decreases the more non-linear that model becomes in its parameters (McKay et al., 1979; Helton and Davis, 2003; Helton et al., 2006). Latin Hypercube sampling is excessively explored in this thesis (Chap. 2) covering in detail the range from the univariate distribution case to the LH sampling of Gaussian or transformed Gaussian (an)-isotropic random fields. LH efficiency is further examined via two synthetic case studies (2D isotropic and 3D anisotropic) involving flow and transport in a heterogeneous porous medium illustrating the benefits of Latin hypercube over simple random sampling.

Stratified likelihood sampling. Along the lines of this dissertation two alterations of an innovative stratified sampling method are introduced (Chap. 3), scarcely known in the geostatistical literature, here named "Stratified Likelihood sampling" (SL) and "Minimum Energy sampling" (ME), in which representative realizations are generated by exploring in a systematic way the structure of the multivariate distribution function itself. This alternate approach of generating stratified realizations was first proposed in a conference proceedings paper by Switzer (2000) in the context of Gaussian random fields bringing a different perspective on the generation of representative realizations from a multivariate Gaussian probability distribution. Kyriakidis and Gaganis (2013) enhanced the method in a spatial context further investigating its performance in a hydrogeological case study. The basic idea involves selecting via stratification representative "points" (sample realizations) from a multivariate likelihood model; hence the term stratified likelihood sampling adopted in this work. These points or realizations lie on multidimensional hyper-ellipsoids, whose geometry is dictated by the eigenvectors and eigenvalues of the covariance matrix of the Gaussian random field model linked to the lognormal model. Realizations lying on the same hyper-ellipsoid can be indexed by a common quantile of the chi distribution, which is none other than the Mahalanobis distance between those realizations and the ensemble mean, as well as their likelihood; hence, such realizations are equally likely. Representative sampling from a lognormal random field in this context involves a controlled (stratified) selection of different hyper-ellipsoids and the subsequent controlled (stratified) layout of points or realizations on the surfaces of such hyper-ellipsoids. This different way of looking at representative sampling from a random field model provides a fruitful alternative to marginal stratification involved in LH sampling, and paves the way for exploring connections with the recent comprehensive approach to uncertainty modeling advocated by (Caers, 2011).

Minimum energy sampling. Although SL sampling constitutes an innovative and efficient stratified sampling within the alternate approach of systematically exploring the multivariate distribution, it is exposed to defects related to the selection of points on the surface of a unit (hyper)sphere; the current implementation of SL sampling is approximate to this respect, and therefore sub-optimal. A similar task arises in geostatistics during the selection of line orientations to efficiently span a unit-sphere in the turning bands simulation algorithm in three dimensions (Chilès and Delfiner, 1999). This representative point placement problem is related to the celebrated best packing problem in mathematics (Hardin and Saff, 2004), which is notoriously difficult to solve in a general way. Although quasi-deterministic solutions exist for such point placement within the context of quasi-Monte Carlo methods (Lemieux, 2009), we propose an energy reduction functional, here named minimum energy (ME) sampling for describing point arrangement by quantifying the minimum distance between nearest neighboring points and then proceed by solving a stochastic optimization problem until the system reaches the state of convergence. ME sampling (Sec. 3.3) is evaluated through a two step spatial hydrogeological case study of flow and transport of a pollutant, along the same lines of previously adopted case studies, and is proved to offer a viable alternative to existing stratified sampling methods.

Two step sampling. A shortcoming of stratified sampling methods is that they aim exclusively at stratification of model parameters, e.g., simulated hydraulic conductivity values, without accounting for the sensitivity of the particular model, e.g., flow and transport, to those values. An additional contribution of this thesis lies in an attempt to fill this gap by proposing a two-step implementation of stratified sampling methods (Sec. 3.2). More precisely, locations where uncertainty in model predictions is expected to be highest, or where attribute values are expected to impact significantly model predictions, are first identified as control locations. Stratified sampling is employed to generate representative realizations of attribute values at those locations, and attribute values at all remaining locations are then generated conditional on the previously simulated values on control locations. The selection of control locations could be guided via analytical solutions (when possible) or from a limited set of model evaluations using attribute realizations generated via simple random sampling. The impact of selecting a particular set (location and number) of control locations on the reproduction of statistics of model outputs, is investigated in this work via a synthetic case study involving a flow and transport model over a simple geometrical domain, assuming known boundary and initial conditions. Results showed that careful selection of important sites, in terms of controlling sampling variability of model outputs across multiple simulations, is proved to yield more representative realizations of model outputs over the entire region.

LH sampling on large grids. Already existing literature on LH sampling methods rely on the Cholesky decomposition of a covariance matrix related to all possible pairs of simulation grid nodes. Since Cholesky decomposition is prohibitive for large covariance matrices, it could be argued that all previously proposed LH sampling approaches are considered computationally inapplicable to case studies composed by very large simulation grids, which in many cases can be the context of hydrogeological investigations. Although there exist simulation methods for generating realizations of random fields on large regular grids, these methods have not been employed to date within a Latin hypercube sampling context. Along these lines, this dissertation proposes a novel simulation method for generating Latin hypercube samples from second order stationary Gaussian random fields on very large grids (with millions of nodes). The proposed LH sampling method (Chap. 4) combines an existing simulation method for creating SR samples from second order stationary random fields in large regular grids (Dietrich, 1997), with the LH sampling method of Stein (1987). The performance of the proposed LH sampling method was compared to that of simple random (SR) sampling via two synthetic case studies involving a two dimensional and a three dimensional flow and transport system, assuming a porous medium whose saturated hydraulic conductivity is parameterized in terms of a lognormal random field with a stationary covariance model. Results proved the proposed LH sampling more efficient than SR sampling, thus reducing the time and computer resources required to perform uncertainty analysis in hydrogeological flow and transport problems discretized by very large regular grids. According to an extensive study of the literature, it is the first time LH sampling is applied for the simulation of random fields in large grids; hence, it is the first time efficient geostatistical simulation can be preformed at models with detailed discretization of the spatially distributed parameters.

Conditional LH sampling on large grids. Last but not least, we explored conditional efficient sampling in terms of applying the LH simulation method on very large grids to a conditional hydrogeological setting reproducing apart from the variogram model of the lognormal random field, any measurements of conductivity at their sample locations. More precisely Chap. 5 thoroughly examines the theoretical context of conditional sampling, and proceeds in proposing a novel combination of LH sampling on very large grids, and conditional sampling via indirect simulation of the Kriging error; the latter overtaking Cholesky decomposition of the covariance matrix, thus also applicable on very large application domains. The proposed simulation method is compared to conditional SR sampling in a synthetic hydrogeological case study of flow and transport on a three dimensional very large simulation domain. Once again, results indicate that the proposed conditional LH simulation method can reproduce statistics of the conductivity and concentration fields with smaller sampling variability than SR sampling for the same number of realizations; thus reducing the computational cost for an uncertainty analysis related to the dependency of variability of model inputs and model outputs.

Future work. In summary, this work offers novel extensions of stratified sampling in a spatial context. All of the proposed methodological approaches could contribute to a wider application of uncertainty analysis endeavors in environmental impact assessment studies. However, stratified sampling approaches place more confidence than simple random sampling on the parameters (mean vector and covariance matrix) of the random field model, as they aim at generating realizations that closely match those statistics; such hyper-parameters, however, are often highly uncertain. Viewed from this perspective, this work only addressed the propagation of spatial variability of input parameters to model output uncertainty. Future work should involve the systematic propagation of hyper-parameter uncertainty to the realizations of the random field and by extension to the associated model predictions; this could be done, for example, within the context of a sensitivity analysis to the overall attribute mean, as well as to the variogram model type, sill and range. Such uncertainty has often a much more pronounced effect on model predictions uncertainty than spatial variability.

Additionally, the most challenging extension of this work is to address real-life hydrogeological problems consisting of complex and non-linear three dimensional groundwater models with millions of nodes and irregular boundary conditions. The number of Monte-Carlo runs required at these cases, depends on the number of uncertain parameters and on the relative accuracy required for the distribution of model predictions. In the context of sensitivity studies, inverse modelling or Monte-Carlo analyses, the ensuing computational burden is usually overwhelming and computationally impractical. These tough computational constrains have to be relaxed and removed before meaningful stochastic groundwater modeling applications are possible (Li et al., 2003).

Future steps of this thesis will focus on the application of the proposed stratified sampling in three dimensional non-linear settings, provided that the marginals of the multivariate probability distribution of the random field model adopted are not tied to any Gaussian assumption. Most sampling applications in a hydrogeological context to date involve Gaussian-related, e.g., lognormal, random field models. However, real-world applications of uncertainty propagation in the earth sciences, especially hydrogeological and petroleum applications, involve (i) parameters with highly non-Gaussian spatial distributions, e.g., hydraulic conductivity fields exhibiting channel-type patterns, (ii) highly non-linear transfer functions, e.g., models of multiphase flow, and (iii) uncertain reservoir geometry and boundary/initial conditions. To address issue (i), a multiple-point geostatistical simulation will be adopted, similar to Mariethoz and Caers (2015), which is capable of generating non-Gaussian spatial distributions by borrowing spatial patterns 'learned' from training images. Addressing issues (ii) and (iii) will be accomplished by using model evaluations in an iterative search to identify input realizations that do correspond (through the lenses of the particular transfer function) to a representative sample of model responses. Stratified sampling methods advocated in this work are planned be used to frame an efficient set of initial 3D realizations for implementing the iterative investigation.

Bibliography

Alabert, F. (1987), 'The practice of fast conditional simulations through the lu decomposition of the covariance matrix', *Mathematical Geology* **19**(5).

Anderson, E. and Hattis, D. (1999), 'Foundations: A. uncertainty and variability', *Risk Analysis* **19**(1).

Ang, A. H.-S. and Tang, W. (1990), *Probability Concepts in Engineering Planning and Design. Volume II: Decision, Risk, and reliability*, John Wiley & Sons, New York.

Anselin, L. and Griffith, D. (1988), 'Do spatial effects really matter in regression analysis?', *Papers of the Regional Science Association* **65**, 11–34.

Armstrong, M., Ndiaye, A., Razanatsimba, R. and Galli, A. (2013), 'Scenario reduction applied to geostatistical simulations', *Mathematical Geosciences* **45**(2), 165–182.

Bart, J. (1995), 'Acceptance criteria for using individual-based models to make management decisions', *Ecological Applications* **5**(2), 411 420.

Benke, K., Lowell, K. and Hamilton, A. (2008), 'Parameter uncertainty, sensitivity analysis and prediction error in a water-blance hydrological model', *Mathematical and Computer Modelling* **47**, 11341149.

Bennett, D., Martinez-Meyer, E. and Feddema, J. (2000), The effect of measurement error on habitat utilization models, *in* Banff, ed., 'In: Proceed-ings of the Fourth International Conference on Integrating GIS and Environmental Modeling (GIS/EM4): Problems, Prospects, and Research Needs, September 28.', Alberta, Canada.

Beven, K. (1989), 'Changing ideas in hydrology-the case of physically based models', *Journal of Hydrology* **105**, 157–172.

Bilcke, J., Beutels, P., Brisson, M. and Jit, M. (2011), Accounting for methodological, structural, and parameter uncertainty in decision-analytic models, Technical Report Medical Decision Making.

Borovkov, K. (1994), 'On simulation of random vectors with given densities in regions and on their boundaries', *Journal of Applied Probability* **31**, 205–220.

Bourgault, G. (2012), 'On the likelihood and fluctuations of Gaussian realizations', *Mathematical Geosciences* **44**, 1005–1037.

Bowman, J. (n.d.), 'Code distribution for minimum energy points on hyperspere', `http://www.mathworks.com/matlabcentral/newsreader/view_thread/21747`.

Caers, j. (2001), 'Geostatistical reservoir modelling using statistical pattern recognition', *Journal of Petroleum Science and Engineering* **29**(3), 177–188.

Caers, J. (2011), *Modeling Uncertainty in the Earth Sciences*, John Wiley & Sons, New York.

Canters, F., De Genst, W. and Dufourmont, H. (2002), 'Assessing effects of input uncertainty in structural landscape classification', *International Journal of Geographical Information Science* **16**(2), 129–149.

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

Cormack, R. (1988), 'Statistical challenges in the environmental sciences: a personal view', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **151**, 201–210.

Cressie, N. (1993), *Statistics for Spatial Data*, revised ed. edn, John Wiley & Sons, New York.

Crosetto, M. and Tarantola, S. (2001), 'Uncertainty and sensitivity analysis: tools for gis-based model implementation', *International Journal of Geographical Information Science* **15**(5), 415437.

Daley, J. (1991), *Atmospheric Data Analysis*, Cambridge University Press, Cambridge, UK.

Darcy, H. (1856), *Les Fontaines Publiques de la Ville de Dijon*, Victor Dalmont, Paris, France.

Davis, J. (1986), *Statistics and Data Analysis in Geology*, 2nd edition edn, John Wiley & Sons, New York.

Davis, M. (1987), 'Production of conditional simulations via the lu triangular decomposition of the covariance matrix', *Mathematical Geology* **19**(2).

Davis, P. (1994), *Circulant Matrices.Second Edition*, New York, NY: Chelsea Publishing.

De Genst, W., Canters, F. and Gulinck, H. (2001), 'Uncertainty modeling in buffer operations applied to connectivity analysis', *Transactions in GIS* **5**(4), 305–326.

de Marsily, G. (1986), *Quantitative Hydrogeology*, Academic Press, San Diego, CA.

Deutsch, C. and Journel, A. (1998), *GSLIB: Geostatistical Software Library and User's Guide*, 2nd edn, Oxford University Press, New York.

Dietrich, C. (1997), 'Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix', *Journal on Scientific Computing* **18**, 1088–1107.

Dowing, D., Gardner, R. and Hoffman, F. (1985), 'An examination of response-surface methodologies for uncertainty analysis in assessment models', *Technometrics* **27**(2), 151–163.

Dunning, J., Stewart, D., Danielson, B., Noon, B., Root, T., Lamberson, R. and Stevens, E. (1995), 'Spatially explicit population models: current forms and future uses', *Ecological Applications* **5**, 3 – 11.

Elith, J., Burgman, M. and Regan, H. (2002), 'Mapping epistemic uncertainties and vague concepts in predictions of species distri-bution', *Ecological Modelling* **157**, 313329.

Emery, W. and Thomson, R. (1997), *Data Analysis Methods in Physical Oceanography*, Pergamon, Oxford, UK.

Euler, L. (1782), 'Recherches sur une nouvelle espece de quarres magiques', *Verhandelingen uitgegeven door het zeeuwsch Genootschap der Wetenschappen te Vlissingen 9* pp. 85–239.

Fahrig, L. and Merriam, G. (1985), 'Habitat patch connectivity and population survival', *Ecology* **66**, 1762–1768.

Fisher, P. (1999), 'Models of uncertainty in spatial data', *In Longley P, Goodchild M F, Maguire D J, and Rhind D W (eds) Geographical Information Systems: Principles, Techniques, Management and Applications (Volume 1)* **1**, 191205.

Frey, C. (1992), Quantitative analysis of uncertainty and variability in environmental policy making, Technical report, Department of Civil Engineering, North Carolina State University.

Frey, C. and Burmaster, D. (1997), 'Methods for characterizing variability and uncertainty: Comparison of bootstrap simulation and likelihood-based approaches', *Risk Analysis* **19**, 109–130.

Frey, C. and Cullen, A. (1999), *Probabilistic Techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*, Springer, New York.

Gaganis, P. and Smith, L. (2001), 'Bayesian approach for evaluating the effect of model error on the predictions of groundwater models', *Water Resour Research* **37**, 2309–2322.

Gaganis, P. and Smith, L. (2006), 'Evaluation of the uncertainty of groundwater model predictions associated with conceptual errors: A per-datum approach to model calibration', *Advances in Water Resources* **29**, 503–514.

Gaganis, P. and Smith, L. (2008), 'Accounting for model error in risk assessments: Alternatives to adopting a bias towards conservative risk estimates in decision models', *Advances in Water Resources* **31**, 1074–1086.

Goodchild, M. (2008), *Foreword; Imprecision and Spatial Uncertainty; Spatial Data Analysis*, Springer, New York, chapter 8, pp. 480–483.

Goovaerts, P. (1997), *geostatistics for natural resources evaluation*, Oxford University Press, New York.

Goovaerts, P. (2001), 'Geostatistical modelling of uncertainty in soil science', *Geoderma* **103**, 326.

Gutjahr, A. and Bras, R. L. (1993), 'Spatial variability and its impact on waste disposal', *Reliability Engineering and System Safety* **42**, 293–316.

Haan, C. (1989), 'Parametric uncertainty in hydrology modelling', *Transactions of the ASABE* **32(1)**, 0137–0146.

Hansen, A., Rotella, J., Kraska, M. and Brown, D. (1999), 'Dynamic habitat and population analysis: An approach to resolve the biodiversity manager's dilemma', *Ecological Applications* **9**(4), 1459–1476.

Hardin, D. and Saff, E. (2004), 'Discretizing manifolds via minimum energy points', *Notices of the American Mathematical Society* **51**, 1186–1194.

He, X., Sonnenborg, T., Jorgensen, F., Hyer, A.-S., Moller, R. and Jensen, K. (2013), 'Analyzing the effects of geological and parameter uncertainty on prediction of groundwater head and travel time', *Hydrology and Earth System Sciences* **17**, 3245–3260.

Helton, J. C. and Davis, F. J. (2002), 'Illustration of sampling-based methods for uncertainty and sensitivity analysis', *Risk Analysis* **22**(3), 591–622.

Helton, J. C. and Davis, F. J. (2003), 'Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems', *Reliability Engineering and System Safety* **81**(1), 23–69.

Helton, J. C., Johnson, J. D., Salaberry, C. J. and Storlie, C. B. (2006), 'Survey of sampling based methods for uncertainty and sensitivity analysis', *Reliability Engineering and System Safety* **91**.

Hoef, J., Cressie, N., Fisher, R. and Case, T. (2001), Uncertainty and spatial linear models for ecological data, *in* 'In Hunsaker, C.T., Goodchild, M.F.. Friedl, M.A., Case, T.J. (eds.), Spatial Uncertainty for Ecology: Implications for Remote Sensing and GIS Applications', Springer-Verlag, New York, pp. 214–237.

Hoffman, F., Gardner, R. and Eckerman, K. (1982), Variability in dose estimation associated with the food chain transport and ingestion of selected radionuclides, Technical report.

Hunsaker, C., Goodchild, M., Friedl, M. and Case, T. (2001), *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications*, Springer, New York.

Iman, R. L. and Conover, W. J. (1982), ‘A distribution-free approach to inducing rank correlation among input variables’, *Communications in Statistics, Part B. Simulation and Computation* **11**(3), 311–334.

IPCC (2000), Good practice guidance and uncertainty management in national greenhouse gas inventories, Technical report.

Jager, H. and King, A. (2004), ‘Spatial uncertainty analysis and ecological models’, *Ecosystems* **7**, 841847.

Jager, H., King, A., Schumaker, N., Ashwood, T. and Jackson, B. (2005), ‘Spatial uncertainty analysis of population models’, *Ecological Modelling* **185**, 13–27.

Jakab, N. (2016), ‘Uncertainty assessment based on scenarios derived from static connectivity metrics’, *Open Geosciences* **8**(1), 799–807.

Jang, C.-S. and Liu, C.-W. (2004), ‘Geostatistical analysis and conditional simulation for estimating the spatial variability of hydraulic conductivity in the choushui river alluvial fan, taiwan’, *Hydrological Processes* **28**, 1333–1350.

Johnson, M. E. (1987), *Multivariate Statistical Simulation*, John Wiley & Sons.

Johnson, R. A. and Wichern, D. W. (1998), *Applied Multivate Statistical Analysis*, 4 edn, Prentice Hall, Upper Saddle River, New Jersey.

Journel, A. (1974), ‘Geostatistics for conditional simulation of ore bodies’, *Economic Geology* **69**, 673–687.

Journel, A. and Huijbregts, C. (1978), *Mining Geostatistics*, Academic Press, London.

Kamali, M., Omidvar, A. and Kazemzadeh, E. (2013), ‘3d geostatistical modeling and uncertainty analysis in a carbonate reservoir, sw iran’, *Journal of Geological Research* **12**, 36–48.

Kendall, M. (1945), *The Advanced Theory of Statistics*, Vol. I, 2nd edn, Charles Griffin & Company, London.

Kroese, D., Taimre, T. and Botev, Z. (2011), *Handbook of Monte Carlo Methods*, John Wiley & Sons, Hoboken.

Kyriakidis, P. (2001), Geostatistical models of uncertainty for spatial data, *in* ‘In: Hunsaker C.T., Hunsaker C.T. (Eds.), Spatial uncertainty in ecology: implications for remote sensing and GIS applications’, Springer, New York.

Kyriakidis, P. and Gaganis, P. (2013), ‘Efficient simulation of (log)normal random fields for hydrogeological applications’, *Mathematical Geosciences* **45**(5), 531–556.

Lemieux, C. (2009), *Monte Carlo and Quasi-Monte Carlo Sampling*, Springer, New York.

Li, S., McLaughlin, D. and Liao, H. (2003), ‘A computationally practical method for stochastic groundwater modeling’, *Advances in Water Resources* **26**(11), 1137–1148.

Lin, Y.-P., Chen, Y.-W., Chang, L.-C., Yeh, M.-S., Huang, G.-H. and Petway, J. (2017), 'Groundwater simulations and uncertainty analysis using modflow and geostatistical approach with conditioning multi-aquifer spatial covariance.', *Water* **9**(3), 1–17.

Luis, S. and McLaughlin, D. (1992), 'A stochastic approach to model validation', *Advances in water resources* **15**(1), 15–32.

Mariethoz, G. and Caers, J. (2015), *Multiple-Point Geostatistics: Stochastic Modeling with Training Images*, Wiley Blackwell, Chichester, UK.

McCarthy, M., Burgman, M. and Ferson, S. (1995), 'Sensitivity analysis for models of population viability', *Biological Conservation* **73**, 93–100.

McDonald, M. and Harbaugh, A. (1988), A modular three-dimensional finite difference ground-water flow model, Technical Report Techniques of Water-Resources Investigations, Book 6: Modeling Techniques, U.S. Geological Survey.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979), 'A comparison of three methods for selecting values of input variables in the analysis of output from a computer code', *Technometrics* **21**(2), 239–245.

Minasny, B. and McBratney, A. (2006), 'A conditioned latin hypercube method for sampling in the presence of ancillary information', *Computers and Geosciences* **32**, 1378–1388.

Morgan, M. G. and Henrion, M. (1990), *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge University Press*, Cambridge University Press, Cambridge.

National US Research Council (1997), Rediscovering geography, Technical report, National Academy Press, Washington, D.C.

National US Research Council (1998a), Hydrologic sciences: Taking stock and looking ahead, Technical report, National Academy Press, Washington, D.C.

National US Research Council (1998b), *The Atmospheric Sciences Entering the Twenty-First Century*, Washington, D.C.

Nicholas, J. and White, D. (2001), *Traceable Temperatures: An Introduction to Temperature Measurement and Calibration*, John Wiley & Sons, New York.

Nilsson, B., Hojberg, A., Refsgaard, J. and Troldborg, L. (2006), 'Uncertainty in geological and hydrogeological data', *Hydrological Earth Systems Science* **3**, 26752706.

Olsson, A. and Sandberg, G. (2002), 'Latin hypercube sampling for stochastic finite element analysis', *Journal of Engineering. Mechanics* **128**, 121125.

Oosterbaan, R., Sharma, D., Singh, K. and Rao, K. (1990), Crop production and soil salinity: evaluation of field data from india by segmented linear regression, *in* 'In: Proceedings of the Symposium on Land Drainage for Salinity Control inArid and

Semi-Arid Regions, February 25th to March 2nd', Vol. 3, Cairo, Egypt, pp. 373 – 383.

Paramelle, L. (1856), *Lart de Decouvrir les Sources*, Librairie Polytechnique, Paris, France.

Pebesma, E. J. and Heuvelink, G. B. M. (1999), 'Latin hypercube sampling of gaussian random fields', *Technometrics* **41**, 303–312.

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

Ripley, B. (1981), *Spatial Statistics*, John Wiley & Sons, New York.

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

Shirmohammadi, A., Chaubey, I., Harmel, R., Bosch, D., Muoz-Carpena, R., Dharmasri, C., Sexton, A., Arabi, M., Wolfe, M., Frankenberger, J., Graff, C. and Sohrabi, T. (2006), 'Uncertainty in tmdl models', *Transactions of the ASABE* **494**, 1033 1049.

Stein, M. (1987), 'Large sample properties of simulations using latin hypercube sampling, Technometrics, 29(2)', *Technometrics* **29**(2), 143–151.

Sudicky, E., Illman, W., Goltz, I., Adams, J. and McLaren, R. (2010), 'Heterogeneity in hydraulic conductivity and its role on the macroscale transport of a solute plume: From measurements to a practical application of stochastic flow and transport theory', *Water Resources Research* **46, W01508**.

Switzer, P. (2000), Multiple simulation of spatial fields, *in* 'in: Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, G.B.M. Heuvelink, and M.J.P.M. Lemmens', Coronet Books Inc., pp. 629–635.

Tarantola, A. (2005), Inverse problem theory and methods for model parameter estimation, Technical Report Society for Industrial and Applied Mathematics, Philadelphia.

Urban, N. and Fricker, T. (2010), 'A comparison of latin hypercube and grid ensemble designs for the multivariate emulation of an earth system model', *Computer Geosciences* **36**, 746755.

Vrugt, J., Stauffer, P., Whling, T., Robinson, B. and Vesselinov, V. (2008), 'Inverse modeling of subsurface flow and transport properties: A review with new developments', *Vadose Zone Journal* **7**(2), 843864.

Wagener, T. and Gupta, H. (2005), 'Model identification for hydrological forecasting under uncertainty', *Stochastic Environmental Research and Risk Assessment* **19**, 378–387.

Ying, Z. (2001), Specification of variogram structures with geometric anisotropy, Technical Report 14th Annual Report of the Stanford Center for Reservoir Forecasting, Stanford University.

Zhang, J. and Goodchild, M. F. (2002), *Uncertainty in Geographical Information*, Taylor and Francis, London.

Zhang, Y. and Pinder, G. (2003), 'Latin hypercube lattice sample selection strategy for correlated random hydraulic conductivity fields', *Water Resources Research* **39**(8), 255–269.

Zheng, C. (1990), MT3D, a modular three-dimensional transport model for simulation of advection, dispersion and chemical reactions of contaminants in groundwater systems, Technical Report Report to the Kerr Environmental Research Laboratory, U.S. Environmental Protection Agency.

Zhou, H., Gomez-Hernandez, J. and Li, L. (2014), 'Inverse methods in hydrogeology: Evolution and recent trends', *Advances in Water Resources* **63**, 22–37.

Zimmerman, D. (1989), 'Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models', *Journal of Statistical Computing and Simulation* **32**, 1–15.

Zuur, G., Ieno, E. N. and Smith, G. M. (2007), *Analysing Ecological Data*, Springer, New York.