Three-Dimensional Non-Multi-Gaussian Simulation of Hydraulic Conductivity Including Multiple Types of Information

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Bo Xiao

aus Nanchang, VR China

Tübingen

2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	12.05.2021
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	PD Dr.-Ing. Claus P. Haslauer
2. Berichterstatter:	Prof. Dr.-Ing. Olaf A. Cirpka

For my grandma and parents

Abstract

Standard geostatistical methods in hydrogeology assume a multi-Gaussian distribution of the log-hydraulic conductivity (K), implying that intermediate values are well connected, embedding isolated zones of high and low values. In this thesis, two datasets of hydraulic conductivity K from the MAcroDispersion Experiment (MADE) site in Columbus, Mississippi, are analyzed, one measured by direct-push injection-logging (DPIL) at 31,123 observation points in 58 vertical profiles and the other by flowmeter profiling at 2611 observation points in 67 wells. The analysis is performed using copula techniques that do not rely on the assumption of multivariate Gaussianity and provide a means to characterize differing degrees of spatial dependence in different quantiles of the K distribution. This characterization provides better insights into the similarities and differences between the two datasets. In addition to the marginal distributions and two-point geostatistical measures, copula-based bivariate rank correlation and asymmetry measures are analyzed and compared. Furthermore, the parameter estimates obtained by likelihood estimation using n-point theoretical models are analyzed. This analysis confirms the similarity of the spatial dependence of K between the two datasets in terms of their marginal distributions and bivariate measures, particularly in the vertical direction. Clear indications of non-multi-Gaussian spatial dependence structures of K are found at this site. The estimation of the K distribution can be improved by taking into account either non-Gaussianity or a censoring threshold, which are expected to lead to a more realistic description of processes that depend on K.

A framework to generate K-field is used that allows non-multi-Gaussian dependence using the multi-objective phase-annealing (PA) method. One objective is to mimic the "asymmetry" of the measured K that indicates the degree of non-Gaussianity. The K-field at the MADE site is mimicked using both DPIL and flowmeter datasets for conditioning. The differences in data quality between the datasets are considered. As the mean and variance of the two datasets differ, the K fields are conditioned on the measured values of the flowmeter dataset and the order within the DPIL dataset. The degree of non-Gaussianity is quantified by the asymmetry of the copula, which is accounted for in the three-dimensional conditioning procedure using the spectral phase-annealing method. The impacts of including as much information as possible in the conditioning procedure on key solute-transport characteristics are analyzed using the comparison between the non-multi-Gaussian method with multi-Gaussian geostatistical approaches. As a transport

metric, the one- and two-particle spatial moments of solute plumes and the associated dispersivities resulting from particle-tracking random-walk simulations are considered. The non-multi-Gaussian models generate preferential flow paths that lead to a stronger correlation of velocity at large separation distances and consequently larger dispersivities in comparison to the (quasi) multi-Gaussian models. A better match between modeled and measured solute transport behavior is obtained when asymmetry is included in the geostatistical model for K.

Kurzfassung

In den herkömmlichen, geostatistischen Methoden der Hydrogeologie wird die hydraulische Leitfähigkeit (K) als log-normalverteilte Zufallsvariable angenommen. Dabei wird implizit vorausgesetzt, dass Zonen mit niedrigen und hohen Werten isolierte Einschlüsse darstellen, wohingegen mittlere Werte gut verbunden sind. In dieser Dissertation werden zwei K Datensätze des MAcroDispersion Experiment (MADE) Feldstandortes analysiert. Der erste stammt von Direct-Push Injection-Logging (DPIL) an 31.123 Beobachtungspunkten in 58 vertikalen Profilen, der zweite von Durchflussmessern (Flowmeter) an 2611 Beobachtungspunkten in 67 Beobachtungsbrunnen. Die hier angewandte Analyse basiert auf der Copula-Methode. Diese beruht nicht auf der Annahme einer multivariaten Normalverteilung. Die Copula-Methode ermöglicht eine differenzierte Analyse der räumlichen Abhängigkeit in unterschiedlichen Quantilen der K-Verteilung. Ähnlichkeiten und Unterschiede der beiden Datensätze werden tiefgehend analysiert. Neben den Randverteilungen und den Variogramm-basierten, bivariaten, geostatistischen Kennwerten werden bivariate Copula-Rangkorrelationen und Asymmetrie analysiert und verglichen. Darüber hinaus werden die Parameterschätzungen von theoretischen n-Punkt Copula-Modellen analysiert. Diese Analyse bestätigt die Ähnlichkeit der räumlichen Abhängigkeit von K zwischen den beiden Daten hinsichtlich ihrer Randverteilungen und der bivariaten Maße, insbesondere in vertikaler Richtung. Die Analyse zeigt das Vorhandensein multivariater, nicht normalverteilter räumlicher Abhängigkeitsstrukturen für K. Indem man entweder eine nichtgaußförmige Abhängigkeit oder eine Zensurschwelle berücksichtigt, kann die Abschätzung der K-Verteilung verbessert werden. Dadurch wird eine realistischere Beschreibung der von K abhängigen Prozesse erreicht.

In dieser Arbeit wird ein methodischer Ansatz verwendet, der K-Felder mit multivariter, nicht normalverteilten Abhängigkeiten auf der Basis von multikriteriellen Temperphasen (Phase Annealing, PA) simuliert. Ziel ist es die Copula Asymmetrie, die von K-Messungen berechnet wird, als ein Maß der nichtgaußförmigen Abhängigkeit als Kriterium in PA einzuschließen. Das K-Feld am MADE-Feldstandort wird nachgeahmt, indem sowohl der DPIL als auch der Durchflussmesser Datensatz für die konditionale Simulation verwendet werden. Die Unterschiede in der Datenqualität zwischen den Datensätze werden berücksichtigt. Weil sich der Mittelwert und die Varianz der beiden Datensätzen unterscheiden, werden die Messwerte des Durchflussmesser-Datensatzes und die Reihenfolge innerhalb des DPIL-Datensatzes im Simulationsverfahren verwendet. Der

Grad der nichtgaußförmigen Abhängigkeit wird durch die Copula-Asymmetrie quantifiziert, die in der dreidimensionalen Konditionierung unter Verwendung der spektralen PA berücksichtigt wird. Weiter wird analysiert, welchen Einfluss das Einbeziehen möglichst vieler Informationen in der konditionalen Simulation auf die wichtigsten Eigenschaften der Tracer-Ausbreitung hat. Dies wird in Bezug auf multivariate, nichtgaußförmige bzw. multivariate gaußförmige geostatistische Ansätze analysiert. Als Transportmetrik werden die räumlichen Ein- und Zweiteilchen-Momente der Tracer-Ausbreitung und der damit verbundenen Dispersivitäten berücksichtigt, die sich aus Random-Walk Simulationen zur Partikelverfolgung ergeben. Multivariate, nichgaußförmige Modelle erzeugen bevorzugte Strömungswege, die stärker mit den Geschwindigkeiten bei großen Trennungsabständen korrelieren und folglich zu größeren Dispersivitäten im Vergleich zu (quasi) mutlivariaten normalverteilten Modellen führen. Dadurch erreicht man eine bessere Übereinstimmung zwischen dem modellierten und gemessenen Ausbreitungsverhalten des Tracers, wenn tiefenabhängige Asymmetrie im geostatistischen Modell für K vorliegt, erreicht.

Acknowledgments

This thesis could not be finished without the help of many people. I would like to thank:

- my supervisor Claus Haslauer for his patient guidance and solid support in my research and life;
- András Bárdossy for the sharing of his unlimited knowledge and intelligent ideas on geostatistics;
- Olaf Cirpka for the sharing of his magnificent knowledge on stochastical hydrogeology;
- Geoff Bohling and Gaisheng Liu, who help me to understand the magic MADE site;
- Sebastian Hörning for the kindly discussion about the phase-annealing;
- Monika Jekelius, Wolfgang Bott, Iris Dreher and Astrid Lemp for their kindly help;
- Research Training Group "Integrated Hydrosystem Modelling" (GRK 1829) for the support;
- all colleagues in the Keplerstr and the center for applied geoscience;
- Mengyun Huang, who helped me to get through tough winter times;
- Jie Ren for her technical support;
- Peijia Ku, who shared a lot of her experience to help me to finish my work;
- my grandma, without her I would not be who I am today;
- my parents for their unconditional support and love.

Contents

Chapter 1

Introduction and Motivation

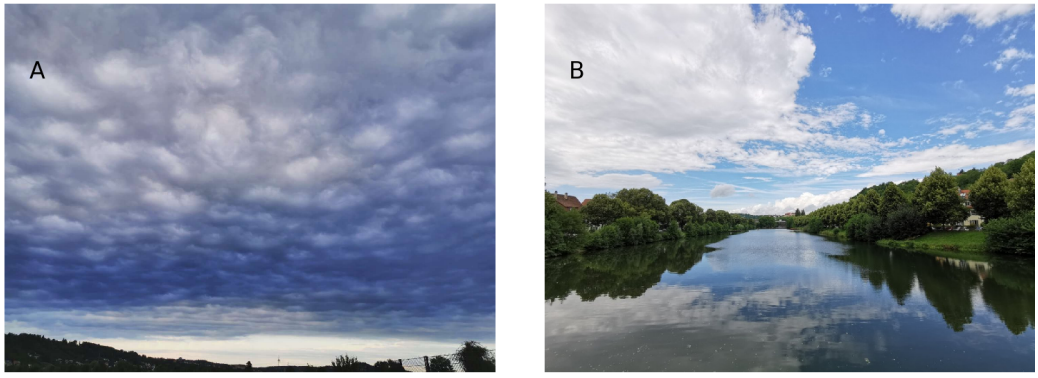

Figure 1.1: Cloud distributions with a different spatial variability. Photos were taken in the Tübingen area, Germany, by the author.

1.1 Background and Motivation

The understanding of the temporal and/or spatial distributions of key parameters of a system on different scales is an important goal for a wide range of scientific and engineering problems. This understanding is not only helpful for the analysis of the system itself but also meaningful for prediction and decision making. Examples of spatiotemporal variables are the distribution of precipitation (Haberlandt, 2007), the distribution of disease cases (Giorgi *et al.*, 2018), the distribution of active regions in the human brain (Ye *et al.*, 2011), etc. The understanding of spatial distributions of important variables is also an urgent task in environmental research. As an example, Figures 1.1A and B show cloud distributions in Tübingen area, Germany, with different spatial variabilities. More observations or a detailed mathematical/physical model are

required to model the variability of a heterogeneous system than to model the variability of a homogeneous system.

The most important parameter in hydrogeology is the hydraulic conductivity (K). The spatial arrangement of K determines the groundwater velocity (Darcy, 1856), and thus advective-dispersive solute transport, for given hydraulic boundary conditions. It is well known that the hydraulic parameters of aquifers vary over orders of magnitude in space. If the structure of K was perfectly known, estimates of solute concentrations over time would be much improved. However, information about the structure of K is sparse because measuring hydraulic conductivity is costly. Different investigation techniques for the estimation of hydraulic conductivity exist, from large-scale multi-well pumping tests, yielding spatially averaged conductivity values, to indirect point estimates inferred from grain-size analyses. K estimations using different methods yield inconsistent results concerning absolute values, are based on different support volumes, and exhibit different data quality. This implies that merging these data in a data-fusion framework needs to address the inconsistencies between the local estimates derived by different investigation methods.

The K-field is inherently uncertain because observations are sparsely distributed and details of the K-field in between remain unresolved. As a consequence, stochastic methods have been widely used for the characterization and interpolation of hydraulic aquifer properties (Gelhar, 1977; Delhomme, 1979). In most standard applications of stochastic subsurface hydrology, the K-field is considered a second-order stationary random space variable that can be described by a stationary log-normal marginal distribution and a stationary variogram (γ). If not stated otherwise, stochastic hydrogeologists assume that the log-conductivity field is multi-Gaussian (Freeze, 1975; Gómez-Hernández and Wen, 1998).

For a solute transport in second-order stationary multi-Gaussian log-conductivity fields with simple mean-flow setups (e.g., uniform-in-the-mean hydraulic gradient), (semi-) analytical solutions have been developed to predict the spatial moments of solute plumes and their uncertainty from geostatistical characteristics of the log-conductivity field (e.g., Gelhar and Axness, 1983; Gelhar, 1986; Rubin, 1990; Dagan *et al.*, 1992; Dentz *et al.*, 2000; Cirpka and Kitanidis, 2000; Renard, 2007; Le Borgne *et al.*, 2008; Rajaram, 2016). Most of these solutions are based on a small-perturbation analysis and are thus restricted to mild heterogeneity. Conditioning on measurements deems the log-conductivity field to become non-stationary, hampering the derivation of closed-form solutions (Dagan, 1982b,a; Graham and McLaughlin, 1989; Zhang and Neuman, 1995). Monte-Carlo (MC) simulations of flow and solute transport are less restrictive concerning the underlying assumptions about the statistical dependence of hydraulic conductivity values at different points, the degree of heterogeneity, or hydraulic boundary conditions. In unconditional simulations, many K fields are generated that are exclusively restricted by the statistical

assumptions but do not honor individual measurements. These unconditional K fields can be conditioned on measurements, which is comparably easy under the assumption of multi-Gaussianity but considerably more complicated when the restriction of Gaussian dependence is relaxed.

A variety of data-driven geostatistical simulation methods are available for the generation of spatially distributed K fields used as input for MC analysis. They use the information from the observations of the primary variable K or additional secondary information for a joint simulation, such as the sequential (Gaussian) simulation (Journel, 1989; Gómez-Hernández *et al.*, 1993), the covariance matrix decomposition using the Cholesky decomposition (Davis, 1987), turning bands method (Matheron, 1973), continuous/discrete spectral method (Mejía and Rodríguez-Iturbe, 1974; Dietrich and Newsam, 1993, 1997) and simulated annealing method (Deutsch and Cockerham, 1994).

The importance of the non-multi-Gaussian spatial dependence structure has been discussed in a series of works since the 1990s (Gómez-Hernández and Wen, 1998; Zinn and Harvey, 2003; Renard, 2007; Kerrou *et al.*, 2008; Meerschaert *et al.*, 2013). Inadequate usage of a multi-Gaussian model can over or underestimate the connectivity of the K-field and further influence the prediction of solute transport behavior. Since then, approaches have been taken to include more information into geostatistic models and move away from the strict limitations of multi-Gaussianity.

Geologic information from lithological characterization has been used in geostatistical approaches using Indicator Kriging and transition probabilities (Fogg, 1996; Carle, 1999). These methods simulate the distributions of the geological units according to the hydrofacies and lithofacies information which is less constrained on the marginal distribution (Fogg *et al.*, 1998; Bianchi *et al.*, 2011; Bianchi and Zheng, 2016; Bianchi and Pedretti, 2017). Another simulation methodology that uses geological expert knowledge is multi-point geostatistics (MPS; (e.g., Strebelle, 2002; Mariethoz *et al.*, 2010; Mariethoz and Caers, 2014; Linde *et al.*, 2015; Pirot *et al.*, 2015; Tahmasebi and Sahimi, 2016b,a; Hansen *et al.*, 2018)), in which geological features of a training image (TI) are associated with the probability that certain multi-point patterns of facies distributions occur. With the help of the development of the neural network and deep learning, Laloy *et al.* (2018) introduced a TI-based spatial generative adversarial neural network method.

The disadvantages of TI-based methods are a) The construction of a TI is difficult, especially when a three-dimensional TI is needed but only one-dimensional line data from boreholes is available. b) Expert knowledge is included in TI. So MPS is not a fully stochastic model. The choice of a good training image is not trivial, as one would want to avoid repetition too much of the typical structures of the TI. It should be possible to simulate features that are outside the observations. Recently, AI-based methods have surfaced in geostatistics (Zhang *et al.*, 2021), that are, like TI-based approaches, difficult

to implement in three dimensions.

In this thesis, datasets from the MAcroDispersion Experiment (MADE) site at the Columbus Air Force Base in Mississippi are used, which has been studied intensively in the past decades. A large number of K measurements and solute concentration measurements offer a unique opportunity for the understanding of the solute transport behavior in a large-scale heterogeneous field site. Compared with other frequently analyzed field sites, such as the Borden (Sudicky, 1986), the Cap Cod (LeBlanc *et al.*, 1991, and others) or the North Bay (Sudicky *et al.*, 2010) sites, the MADE site exhibits large heterogeneity (Mackay *et al.*, 1986, and others). The specific datasets in the present study are inferred from flowmeter tests (Rehfeldt *et al.*, 1992; Boggs *et al.*, 1992) and that derived from high-resolution direct-push injection-logging (DPIL) (Liu *et al.*, 2012; Bohling *et al.*, 2012), in which water is injected over a very short screen, which is advanced by a direct-push rig, and the conductivity is estimated from the ratio of the injection rate over the pressure applied.

Bohling *et al.* (2016) presented a revised calibration of the DPIL data accounting for the insensitivity of DPIL responses to K variations above a certain threshold. When high-K values are encountered, injection-induced pressure changes approach the lower detection limit of the pressure transducer, weakening the quality of the signal. In contrast, the flowmeter measurements are particularly subject to uncertainty in low-K zones due to the low-flow threshold of the impeller flowmeter (Rehfeldt *et al.*, 1992). Measurements above ("on the right side of some threshold", DLR) or below a threshold ("on the left side of the detection limit", DLL), or generally within an interval ("censored measurements") are more uncertain than those that are not censored. Despite this uncertainty, they still provide useful information, namely, that they reside in an interval. How to use this information is a long-standing issue in surface and subsurface hydrology (Gilliom and Helsel, 1986; Liu *et al.*, 1997; Cohn, 2005). The approaches of Bárdossy (2011) and Haslauer *et al.* (2017b) allow the integration of censored measurements into copula-based estimation.

Tracer tests (MADE-1 and MADE-2) performed at the MADE site showed anomalous behavior, indicating that a macroscopic description by the advection-(macro)dispersion equation yields unsatisfactory results because the tracer plumes are skewed (Zheng *et al.*, 2011; Gómez-Hernández *et al.*, 2017). This has fostered the development of alternative effective transport models, such as Rate-Limited Mass Transfer (RLMT) methods (Harvey and Gorelick, 2000), Dual-Domain Mass-Transfer (DDMT) models (Feehley *et al.*, 2000), Continuous-Time Random-Walk (CTRW) methods (Berkowitz and Scher, 1998), Multi-Rate Mass-Transfer (MRMT) models (Guan *et al.*, 2008), time-domain random-walk (TDRW) (Cvetkovic *et al.*, 1996; Fiori *et al.*, 2007) and fractional advection-dispersion equation (fADE) models (Zhang and Benson, 2008). These effective transport models contain additional macroscopic parameters that, unlike macrodispersivities derived by linear theory, are difficult to relate to characteristics of the underlying hydraulic conductivity field and are thus often determined by fitting simulated breakthrough curves to measured

tracer data.

Many studies have been performed at the MADE site to model the anomalous behavior of the field tracer test result. Harvey and Gorelick (2000) used the RLMT model to describe the interaction between the solute with the aquifer material as a linear rate-limited mass transfer. Feehley *et al.* (2000) used the DDMT to model the heterogeneous aquifer as a mobile domain, in which the transport is dominated by advection, and an immobile domain, in which the transport is dominated by diffusion, that are connected by a mass transfer. Julian *et al.* (2001) combined the DDMT model on a multiscale K-field to catch the large and small scale heterogeneity and modeled the MADE-3 (NAT) result. Bowling *et al.* (2006) used the DDMT model on a K-field derived from the direct-current resistivity data. Liu *et al.* (2008) used the DDMT model with parameter assimilation using the ensemble Kalman filter (EnKF). Liu *et al.* (2010) used DDMT with a three-zone K-field based on the DPIL observations to model the MADE-4 tracer test result. Guan *et al.* (2008) used a dual-porosity transport model with an assumption that the injected tracer was trapped hydraulically near the injection site.

Berkowitz and Scher (1998) and Berkowitz *et al.* (2006) used CTRW to model the plume distribution at the MADE site. Salamon *et al.* (2007) showed that the modeling of the K-field can be improved using a hole-effect geostatistical model. Li *et al.* (2011) used a Laplacian-based upscaling technique coupled to a non-uniform coarsening scheme to upscale the K-field, which is simulated using a hole-effect geostatistical model (Salamon *et al.*, 2007). Edery *et al.* (2014) combined CTRW with a particle-visitation weighted histogram to describe the preferential pathways across each domain. Cvetkovic *et al.* (2014) used a TDRW-model combining the self-consistent approximation (SCA) with K fields that are simulated by the multi-indicator method (MIM) and showed that the travel time distribution can not be modeled using the common distributions, e.g., log-normal distribution or the inverse-Gaussian distribution, used for modeling hydrogeological transport. Dentz *et al.* (2020) used a TDRW-model with an upscaled Lagrangian approach to model the mass distribution of the tracer test.

Linde *et al.* (2015) showed the importance of the TI while using MPS to mimic the K structure at the MADE site. Ronayne *et al.* (2010) used an ADE model on a hybrid K model combining three-dimensional lithofacies on a background correlated multivariate Gaussian matrix. Bianchi *et al.* (2011) investigated the connectivity of the K-field using the transition probabilities. Bianchi and Zheng (2016) used an ADE-based model with K fields, which are derived from the lithofacies distribution based-on a transition probabilities model.

Benson *et al.* (2001) used the fADE to model the non-Fickian behavior at the MADE site by matching the order of differentiation of the dispersive derivative to the exponent of the plume growth process. Schumer *et al.* (2003) used a fractal mobile/immobile

model and assumed power-law waiting times in the immobile zone. Zhang and Benson (2008) extended the fADE model in a Lagrangian framework to prevent the explicit definition of the local-scale heterogeneity. Dogan *et al.* (2014) used an ADE-model with a multiscale fractal log-normal K-field generator conditioned on the flowmeter value and using autocorrelation functions of the DPIL observations.

Jankovic *et al.* (2017) and Fiori *et al.* (2017) used an ADE with First Order Approximation (ADE-FOA) and showed that other sources of uncertainty, e.g., the mean velocity, are more important than the K structure. Fiori *et al.* (2019) further confirmed that the non-ergodic effects and uncertainty of parameters, especially the uncertainty of the mean velocity, contribute to a large extent to the bias in solute transport modeling.

Barlebo *et al.* (2004) suggested using the inverse model to match the hydraulic head and concentration measurements instead of the hydraulic conductivity observations. (Fiori *et al.*, 2013) used a local-ADE (LADE) model combining the SCA with K fields that are simulated using MIM and showed the importance of the spatial variability of local advection and conductivity. Dünser and Meyer (2016) used a polar Markovian velocity process (PMVP) to reduce the computational cost of the Monte-Carlo simulation.

The local variability (Salamon *et al.*, 2007; Fiori *et al.*, 2013) and the heterogeneous domain (Julian *et al.*, 2001; Liu *et al.*, 2010) structure are considered as key points for the modeling of the solute transport at the MADE site. To include these two points in the geostatistical simulation, realistic, highly resolved non-Gaussian conductivity fields are simulated and used in spatially resolved transport simulations to improve the understanding of the observed solute transport behavior at this site.

Previous comparisons (Bohling *et al.*, 2012, 2016) concluded that the flowmeter and DPIL K data display similar large-scale patterns and remarkably similar variograms, despite differences between the marginal distributions of the two datasets. These similarities provide compelling evidence that both datasets accurately reflect the two-point autocorrelation structure of the K-field at the MADE site. Given the differences between observed and simulated solute transport at the site (Harvey and Gorelick, 2000; Feehley *et al.*, 2000), the question naturally arises if some important aspect of the spatial arrangement of K is missed if the analysis is constrained to the first two statistical moments and a multivariate normal spatial dependence is assumed.

A certain aspect of the potentially missed or not described spatial structure can be captured by the multivariate copula. Copulas were introduced by Sklar (1959), and have been developed in insurance and financial mathematics (Embrechts *et al.*, 2003). Copulas can describe the properties of real-world data without imposing an assumption of multivariate Gaussianity. Recently, they have been used to analyze hydrological and hydrogeological problems, both in terms of spatial data (Bárdossy, 2006; Bárdossy and Li, 2008;

Marcotte and Gloaguen, 2008; Kazianka and Pilz, 2010, 2011; Gräler, 2014) and time series (De Michele and Salvadori, 2003; Favre *et al.*, 2004; Salvadori and De Michele, 2004; Erhardt *et al.*, 2015a,b).

Spatial copulas offer a way to describe non-multi-Gaussian spatial dependence in which the degree of statistical dependence differs in various quantiles and varies with separation distances (Bárdossy, 2006; Bárdossy and Li, 2008; Guthke and Bárdossy, 2017). By using copula-based correlograms and asymmetry functions (A), the spatial fields of interest (here the conductivity field $K(\mathbf{x})$) can be simulated more realistically than with multi-Gaussian geostatistical methods. In particular, multi-Gaussian fields show strong connectivity of intermediate values that embed inclusions of high and low values due to their maximum entropic character, whereas copula-based approaches accounting for asymmetry allow the description and modeling of spatial fields in which low and/or high values show better connectivity. For example, the homogeneous Borden site has been modeled using non-multi-Gaussian copula models and it was shown that with only ten times increased variance of the dataset, solute transport model results significantly differ between non-multi-Gaussian and multi-Gaussian cases that otherwise (mean and covariance function) are statistically identical (Haslauer *et al.*, 2012). For the simulation of non-multi-Gaussian fields, random mixing has proven to be an effective method (Bárdossy and Hörning, 2016).

1.2 Structure of This Thesis

The goal of this thesis is to analyze and simulate the spatial dependence structure of K at the MADE site using copula-based methods without assuming multivariate Gaussianity. The copula-based bivariate and multivariate geostatistical theory and methodology are reviewed in Chapter 2. Chapter 3 presents how to simulate non-multi-Gaussian K fields by including copula-based measures and Chapter 4 introduces the evaluation of the longitudinal dispersivity using a particle-tracking random-walk simulation. In Chapter 5 the flowmeter and DPIL datasets at the MADE site are compared between each other considering their univariate measures, their empirical copula-based bivariate measures, the estimated parameters of Gaussian and v-copula models fitted to both datasets in high dimensional spaces, including model fits accounting for data censoring. Simulated K-field with different multi-Gaussian and non-multi-Gaussian conditional geostatistical models and the influence of the included information on the solute transport are presented in Chapter 6. Conclusions and an outlook are presented in Chapter 7.

Chapter 2

Copula-Based Geostatistical Theory and Methodology

The content in this chapter contains materials published in "Xiao, B., Haslauer, C., and Bohling, G. (2019). Comparison of multivariate spatial dependence structures of DPIL and flowmeter hydraulic conductivity data sets at the MADE site. Water (MDPI), 11(7), 1420".

Author	**Author position**	**Scientific ideas %**	**Data generation %**	**Analysis & interpretation %**	**Paper writing %**
Bo Xiao	1	70	80	60	75
Claus Haslauer	2	20	20	30	20
Geoffrey Bohling	3	10	0	10	5
Titel of paper:	Comparison of multivariate spatial dependence structures of DPIL and flowmeter hydraulic conductivity data sets at the MADE site.				
Status in publication process:	Published.				

In the world of geostatistics, the variable under investigation (in this thesis: the hydraulic conductivity K) is modeled as a regionalized variable, whose distribution is assumed as one realization of a random variable. A random variable is also named a random field (RF) when it is distributed in Euclidean space. The properties of a random field are described by statistical (ensemble) measures. In this chapter, basic statistical measures and geostatistical measures are introduced in Sections 2.1 and 2.2. Then the multi-Gaussian assumption of variogram-based geostatistics is discussed in Section 2.3. In Section 2.4, the concepts of copula-based geostatistical measures and parameter estimation are presented.

2.1 Basic Statistical Measures

A random field Z can be described using statistical measures, e.g., the distribution function F_Z, the density function f_Z, etc., in the probability space P. The discrete form of the (cumulative) distribution function F_Z (CDF) of a random field Z is defined as

$$F_Z(z) = P[Z \leq z], \tag{2.1}$$

in which the probability P of Z in the interval between $(a,b]$ can be calculated as

$$P[a < Z \leq b] = F_Z(b) - F_Z(a). \tag{2.2}$$

The continuous form of $F_Z(z)$ (Equation 2.1) and the (probability) density function $f_Z(z)$ (PDF) is written as

$$F_Z(z) = \int_{-\infty}^{z} f_Z(t)dt \tag{2.3}$$

and

$$f_Z(z) = \frac{d}{dz}F_Z(z). \tag{2.4}$$

Both CDF and PDF can be estimated empirically, using a parametric distribution or using nonparametric kernel density estimation (Chen, 2017).

The m-th moment $E[Z^m]$ and the m-th central moment $E[(Z-E[Z])^m]$

$$E[Z^m] = \int_{-\infty}^{+\infty} t^m f_Z(t)dt, \tag{2.5}$$

$$E[(Z-E[Z])^m] = \int_{-\infty}^{+\infty} (t-E[Z])^m f_Z(t)dt \tag{2.6}$$

of the marginal distribution of a random field describe the shape of the PDF, in which $E[\cdot]$ is the expected value. The first moment denoted mean value μ, the second central moment denoted variance σ^2, the third- and the fourth- central moments known as skewness $skew(Z)$ and kurtosis $kurt(Z)$ are widely used to describe the shape of a marginal distribution. There are various parametric distribution functions used in statistics, e.g., the normal (Gaussian) distribution (Φ), the uniform distribution, etc., which can be fully characterized by their statistical moments. For example, the uniform distribution in the range $[0,1]$ has the mean $\mu = 0.5$ and variance $\sigma^2 = \frac{1}{12}$.

The multivariate joint distribution function of n random fields $Z_1, ..., Z_n$ is defined as:

$$F_n(z_1, ..., z_n) = F_{Z_1,...,Z_n}(z_1, ..., z_n) = P[Z_1 \leq z_1 \cap ... \cap Z_n \leq z_n]. \tag{2.7}$$

$P(A|B)$ is defined as the conditional probability of event A under the condition that event B is true. The multivariate conditional probability is defined as:

$$P_{Z_n}(Z_n = z_n | Z_1 = z_1 \cap ... \cap Z_{n-1} = z_{n-1}) = \frac{P[Z_1 = z_1 \cap ... \cap Z_n = z_n]}{P[Z_1 = z_1 \cap ... \cap Z_{n-1} = z_{n-1}]}. \quad (2.8)$$

2.2 Variogram Based Geostatistics

The basic assumption of geostatistics is that spatially close measurements of a random space variable are more correlated than measurements separated by a larger distance. This spatial dependence can be described using the (semi-)variogram $\gamma(\mathbf{h})$ under the intrinsic hypothesis assumption (Webster and Oliver, 2008), which relates the average variance of measurements at two locations versus their separation vector $\mathbf{h}$:

$$\begin{aligned} \gamma(\mathbf{h}) &= Cov(\mathbf{0}) - Cov(\mathbf{h}); & (2.9)\\ &= \frac{1}{2}E[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2]; & (2.10)\\ &\approx \frac{1}{2N(\mathbf{h})} \sum_{\mathbf{x}_i - \mathbf{x}_j \approx \mathbf{h}} (z(\mathbf{x}_i) - z(\mathbf{x}_j))^2, & (2.11) \end{aligned}$$

in which $Cov(\mathbf{h})$ is the covariance function for a certain separation vector $\mathbf{h}$ and $N(\mathbf{h})$ is the number of data pairs with coordinate vectors $\mathbf{x}_i$ and $\mathbf{x}_j$.

Various theoretical variogram functions are defined to model the spatial dependence structure of a random field based on the observations, for example, the exponential variogram ($Exp(\mathbf{h})$)

$$\gamma(\mathbf{h}) = Cov(0) \cdot (1 - e^{-\frac{|\mathbf{h}|}{a}}), \quad (2.12)$$

the spherical variogram ($Sph(\mathbf{h})$)

$$\gamma(\mathbf{h}) = \begin{cases} Cov(0) \cdot \left[\frac{3}{2}\frac{|\mathbf{h}|}{a} - \frac{1}{2}(\frac{|\mathbf{h}|}{a})^3\right], & for\, |\mathbf{h}| \leq a \\ Cov(0), & for\, |\mathbf{h}| > a, \end{cases} \quad (2.13)$$

and the Matérn variogram ($Mat(\mathbf{h})$)

$$\gamma(\mathbf{h}) = Cov(0) \cdot \left[1 - \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{|\mathbf{h}|}{a}\right)^{\kappa} B_{\kappa}\left(\frac{|\mathbf{h}|}{a}\right)\right], \quad (2.14)$$

in which a is the range, κ is the smoothness parameter, $\Gamma(\cdot)$ is the gamma function and $B_{\kappa}(\cdot)$ is the κ-order modified Bessel function.

Interpolation and simulation are two main geostatistical applications to estimate the distribution of a random field in space. A geostatistical interpolation estimates the spatial distribution of the conditional mean of a spatial variable based on observations and an (assumed) variogram. Kriging (Delhomme, 1978) is one of the widely used geostatistical interpolation methods, in which the variogram-based estimation variance is minimized while guaranteeing unbiasedness. While the estimated mean is mainly influenced by the measurement values of the closest observations, the estimation variance depends only on the distance to the observation points. However, the Kriged values represent the interpolated random field as a conditional mean based on the observations. The variabilities corresponding to different scales are normally underestimated. Therefore, when the Kriged field is used for a further modeling investigation, the unresolved fine-scale variability can affect the macroscopic properties of the modeling process. So a Monte-Carlo (MC) geostatistical simulation method is preferred to perform a stochastic analysis to include the data uncertainty in the model (Deutsch and Cockerham, 1994). Each simulated field describes one possible realization of the investigated random field. The probability of the spatial distribution of the unknown "Truth" is described using the ensemble properties of the realizations.

Unconditional and conditional simulations are two groups of geostatistical simulation methods according to the included information. An unconditional simulation reproduces the marginal distribution and the spatial dependence structure using the variogram. A conditional simulation meets also the data of the primary or secondary variable at certain locations (Chilès and Delfiner, 2012), which comes at additional computational costs.

Figure 2.1 shows a comparison between the unconditional and the conditional simulation to model the reference "truth" with mean value $\mu = 0$ (Figure 2.1A). The unconditional simulation generates realizations with an equal pixel (point) probability (Figure 2.1B). Each point has the same distribution function and ensemble point statistical measures, e.g., mean, variance, and so on, which equal the univariate measures of the observations. So, the mean value (black dashed line) of all realizations is a straight line centered at zero. This property is changed in the conditional simulation. The simulated values are strongly influenced by the nearby conditional points upon conditioning (Figure 2.1C and D), which means that the point-based probability space of the simulated realizations is drifted and varies in space. Local features based on the corresponding conditional points are generated, although the marginal distribution is not changed. Under these conditions, the ensemble mean values (black dashed line) at the locations close to a conditional point can present the reference "truth" better than at the locations far away from a conditional point. This is important when a physical process, which depends on the distribution of the investigated variable, is dominated by local features of this variable, e.g., the preferential paths in the solute transport modeling. In this case, the important local features of the key parameters are eliminated in a stochastic analysis using an unconditional simulation due to the over-averaging effect.

This thesis focuses mainly on the data-driven geostatistical conditional simulations of K fields based on the K observations. Other methods, like training images (TI) based multiple-point geostatistics (MPS), are not discussed in this thesis.

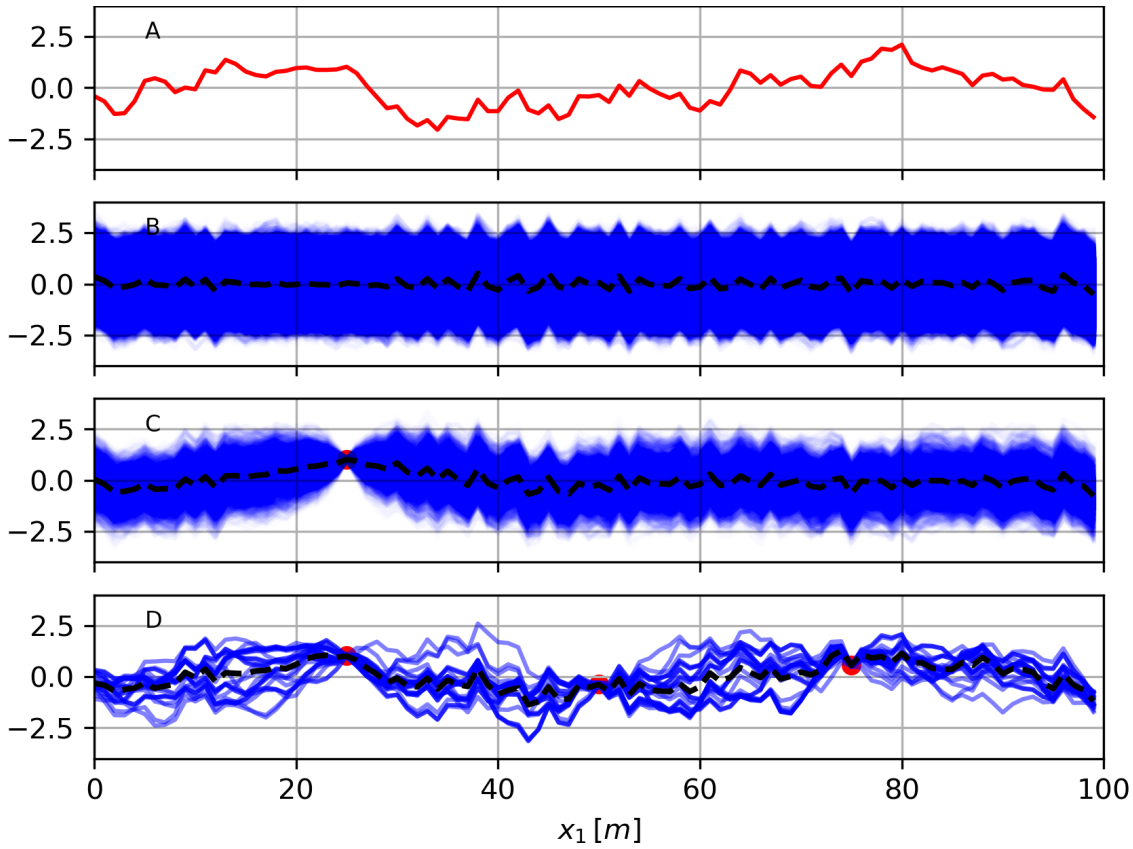

Figure 2.1: A comparison between unconditional and conditional simulations. A) The reference "truth" of a one-dimensional random field. B) Realizations of the unconditional simulation. C) Realizations of the conditional simulation using one conditional point (the red point). D) Realizations of the conditional simulation using three conditional points (red points). The black dashed lines represented the average values of all realizations at every location.

2.3 Gaussianity, Heterogeneity, and Variability

In classical stochastic subsurface hydrology, the distribution of a long-tailed K-field is often assumed to be log-normal. That is, the K-field has a multivariate Gaussian distribution after performing a logarithm transformation. This multi-Gaussian assumption is implicitly included in standard variogram-based geostatistical simulations. The degree

of variability of a K-field is usually described by the variance σ^2 of $\ln(K)$. The variability, which is caused by the departure from multi-Gaussianity, are normally neglected. For given first- and second-central moments, a multi-Gaussian random field has the maximal entropy and the largest disorder in the system (Journel and Deutsch, 1993). Gómez-Hernández and Wen (1998) showed that this multi-Gaussian assumption underestimates the spatial connectivity of extreme high and/or low values, which is important in the modeling of solute transport in heterogeneous systems exhibiting preferential flow paths. In the following parts of this section, some details of the multi-Gaussian assumption are discussed.

2.3.1 Univariate Marginal Distributions

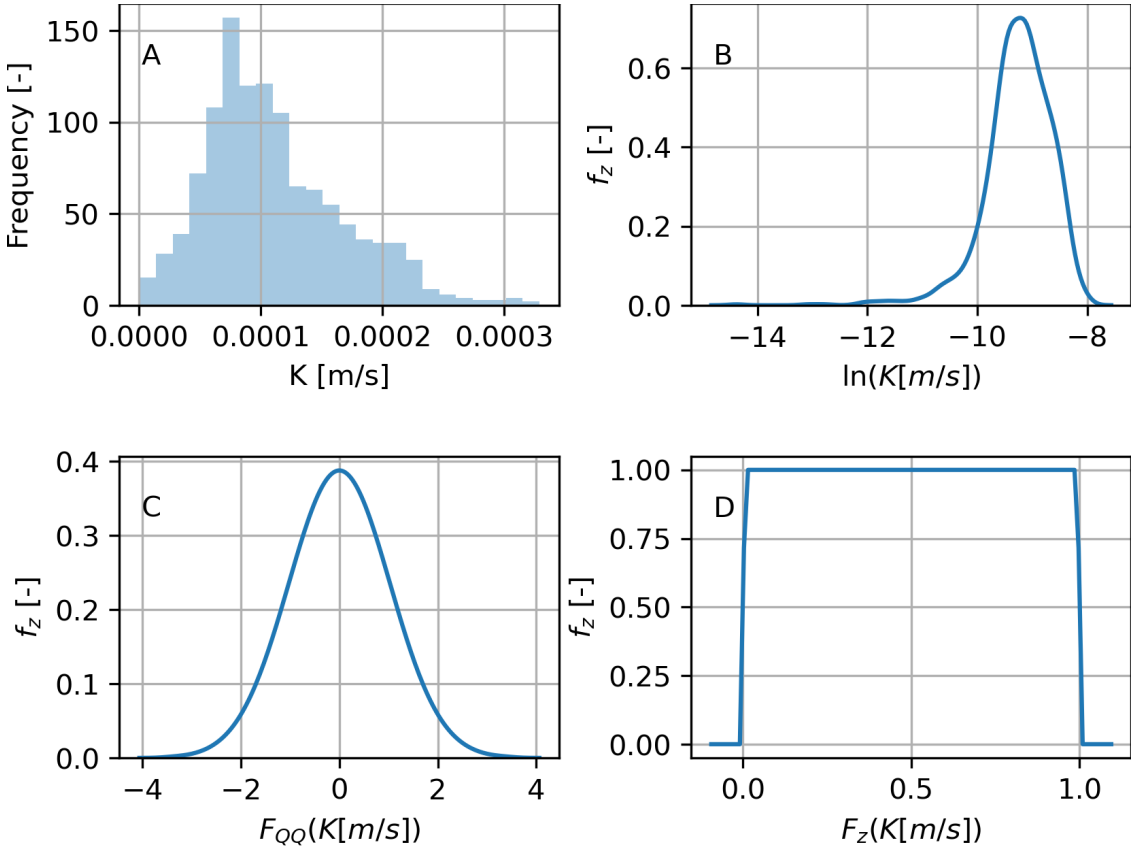

Figure 2.2: Influence of different transformations on the marginal distribution. A) Histogram; B) PDF after the log-transformation C) PDF after the QQ-transformation and D) PDF of the dataset in $F(K)$ space.

Typically, the K-field has a long tail univariate distribution function in a natural aquifer. A convenient way to account for that is to assume that the log-conductivity field is (multi)-Gaussian (Freeze, 1975).

There are, however, also other approaches available to transform the data of a long-tailed distribution to a quasi-Gaussian distribution. The quantile-quantile (QQ) transformation

$$F_{QQ} = F_{2,z}^{-1}(F_{1,z}(z(\mathbf{x}))). \tag{2.15}$$

varies the data of a marginal distribution $F_{1,z}$ to a new marginal distribution $F_{2,z}$. When $F_{2,z}$ is a normal distribution, it is also called the normal-score transformation or Gaussian anamorphosis. Zhou *et al.* (2012) and Schniger *et al.* (2012) use the normal-score transformation to enhance the performance of the ensemble Kalman filter method.

Figure 2.2 shows how the shape of the marginal distribution of a homogeneous dataset (Figure 2.2A with $\sigma^2(\ln(K[m/s]) = 0.39$, (Sudicky, 1986)) is changed after the log-transformation (Figure 2.2B), the QQ-transformation (Figure 2.2C), and transformed in the rank space (F_Z) with Equation 2.4 (Figure 2.2D).

2.3.2 Bivariate Spatial Dependence

The transformed marginal distributions in Section 2.3.1 change how a random field looks like in space. Figure 2.3 shows the pseudocolor plots (A1-D1) and bivariate scatter plots (A2-D2) of the corresponding marginal distribution in Figure 2.2. The transformed marginal distributions determine the empirical variogram model (Equation 2.9) and further influence the fitted theoretical variogram model (Chapter 2.2). However, all transformations in Figure 2.3 are monotonic, which means that the relationship between two values in the rank space is not changed after the transformation. Therefore, the statistical properties are changed only in the value space but are constants in the rank space. The spatial dependence of a random field in the rank space is defined as the underlying spatial dependence structure. The underlying spatial dependence structure of the random fields in Figure 2.3A1-D1 would be identical. In the same way, random fields with an identical variogram can have different underlying spatial dependence structures, i.e., multi-Gaussian or non-multi-Gaussian, which are reflected in a different spatial arrangement of the K values in the rank space.

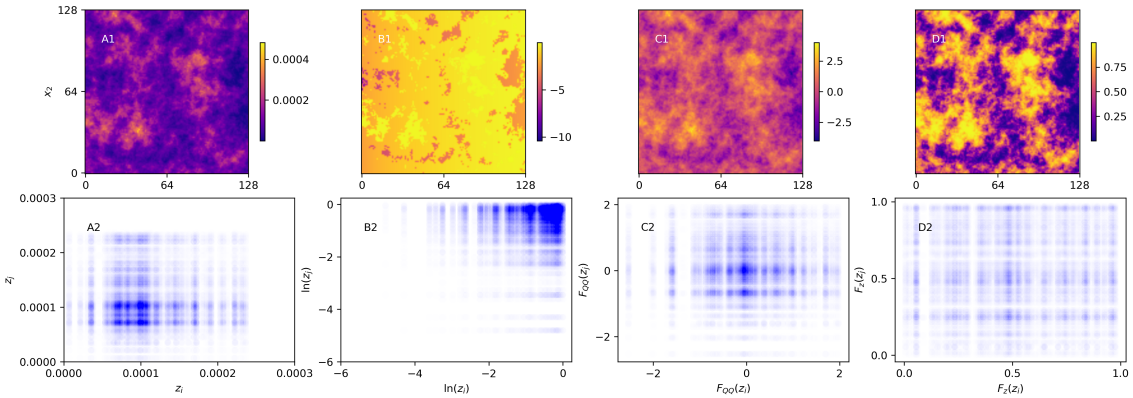

Figure 2.3: Pseudocolor plots of random fields with different marginal distributions (A1-D1), and bivariate plots of same data pairs with different marginal distributions (A2-D2). The corresponding margins from left to right are A) in the original value space, B) in a log-transformed space, C) in a QQ-transformed space D) in distribution function (F_Z) space.

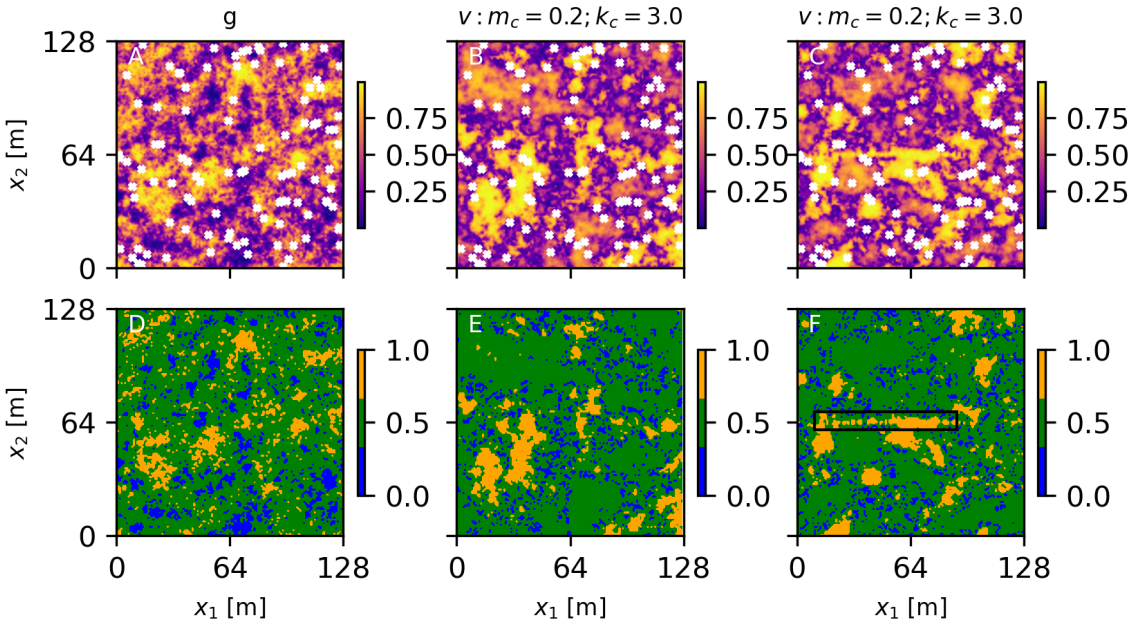

Figure 2.4: Contour plots of random fields with an identical variogram and marginal distribution and A) a multi-Gaussian; B) a non-multi-Gaussian spatial dependence structure C) a non-multi-Gaussian spatial dependence structure with a local feature on the middle of the domain. The white cross signs are the locations with conditional points. The corresponding indicator plots of D) the multi-Gaussian; E) the non-multi-Gaussian random field F) the non-multi-Gaussian spatial dependence structure with a local feature on the middle of the domain (black box).

A useful tool to show the difference between the underlying spatial dependence is the indicator plot. The values are sorted and categorized according to one or more thresholds. The points in the same category are assigned to the same value. Figure 2.4 shows the pseudocolor plots (Figure 2.4A and C) and the indicator plots (Figure 2.4B and D) of two random fields with an identical variogram and marginal distribution but different underlying spatial dependence. The thresholds used to generate the indicator plot are

$$\begin{aligned} 0 < F_Z \leq 0.1 \sim I(F_Z) = &\quad 0; \\ 0.1 < F_Z \leq 0.9 \sim I(F_Z) = &\quad 0.5; \\ 0.9 < F_Z \leq 1.0 \sim I(F_Z) = &\quad 1; \end{aligned} \tag{2.16}$$

and are used as the default thresholds in the following parts of this thesis. It is obvious that in Figure 2.4E the large values ("orange blobs") are more likely to be arranged together than in Figure 2.4D. The solute transport characteristics based on these two K fields would be expected to have different behaviors, which are caused by the degree of non-Gaussianity of the K fields.

It is even more complex when a local feature of a K-field exists (Figures 2.4C and F). For example, two blobs of large values are connected by this high K-values pathway in the black box. The hypothesis is that the solute behavior in this K-field would be strongly influenced by the degree of Gaussianity and this local feature, especially when the injection point is located in this area, which means that an unconditional simulation with a given mean and variogram is not enough to describe the difference between the random fields in Figure 2.4. To model such a K-field, more information must be extracted from the observations and included in a conditional simulation as discussed in Figure 2.1.

2.3.3 Multivariate Gaussianity

N-point Gaussianity of a random field can be tested using n-point correlations in the rank space. Observations are sampled by triangles and squares (Bárdossy and Pegram, 2009). The combination of n-point around the vertex of triangles and squares are classified according to the defined threshold in the rank space between $(0,1)$. Figure 2.5 shows a conceptual plot of this methodology with a threshold of 0.5, in which $z(\mathbf{x})' = 0$ when $z(\mathbf{x}) < 0.5$ and $z(\mathbf{x})' = 1$ when $z(\mathbf{x}) \geq 0.5$. For an n-point correlation with the number of thresholds n_{thres}, there are n_{thres}^n different combinations. For example, triangles have 3-point combinations of $(0,0,0)$, $(1,1,1)$, $(0,0,1)$, $(0,1,0)$, $(1,0,0)$, $(0,1,1)$, $(1,0,1)$, and $(1,1,0)$. Figure 2.6 presents the ratio of the number of each 3- and 4-point combination on the random fields in Figure 2.4A (left column) and B (right column). A different composition can be found on a multi-Gaussian (left column) and on a non-multi-Gaussian (right column) random field. The compositions in Figure 2.6 can be further evaluated

using measures like entropy.

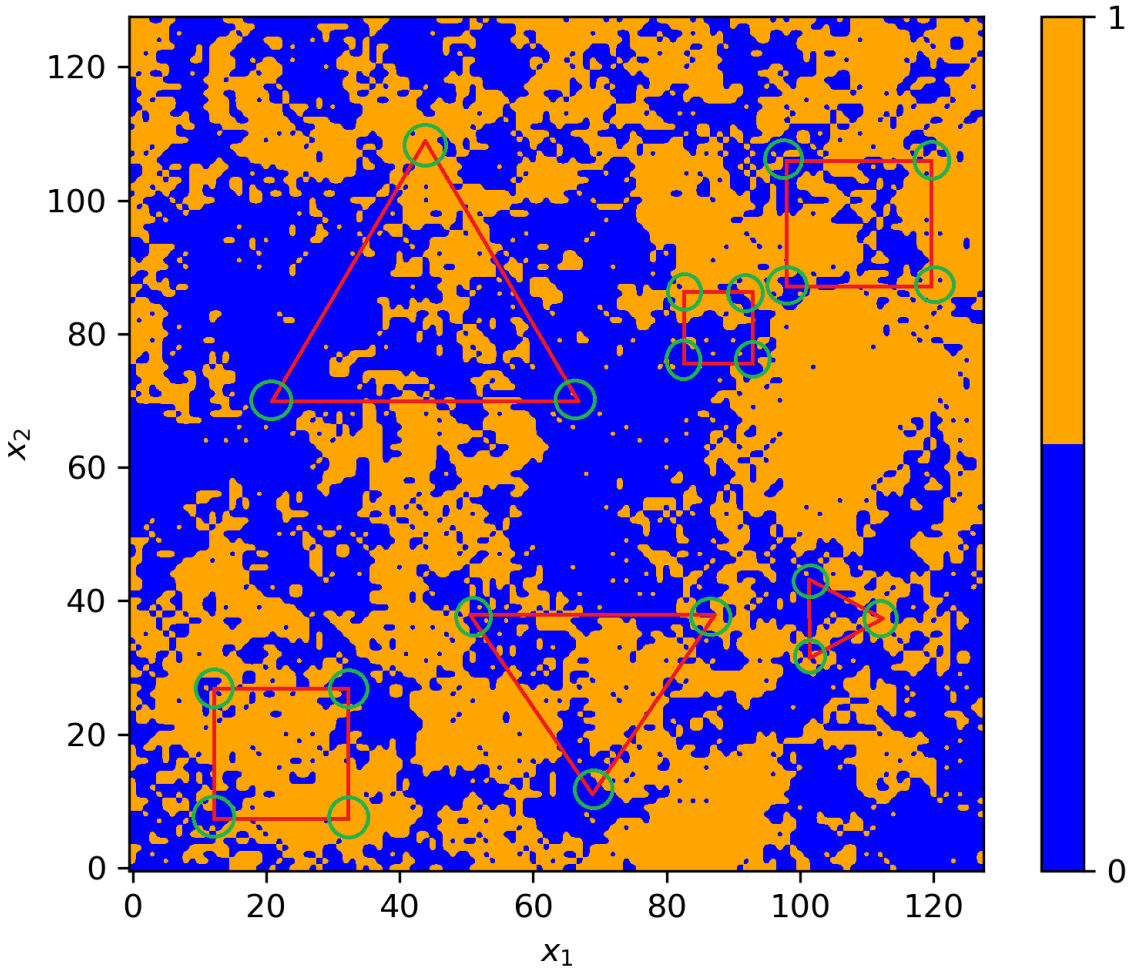

Figure 2.5: Conceptualization of the evaluation of n-point spatial dependence structure. 3-point: red triangles; 4-point: red squares; tolerant range for the point selection: green cycles.

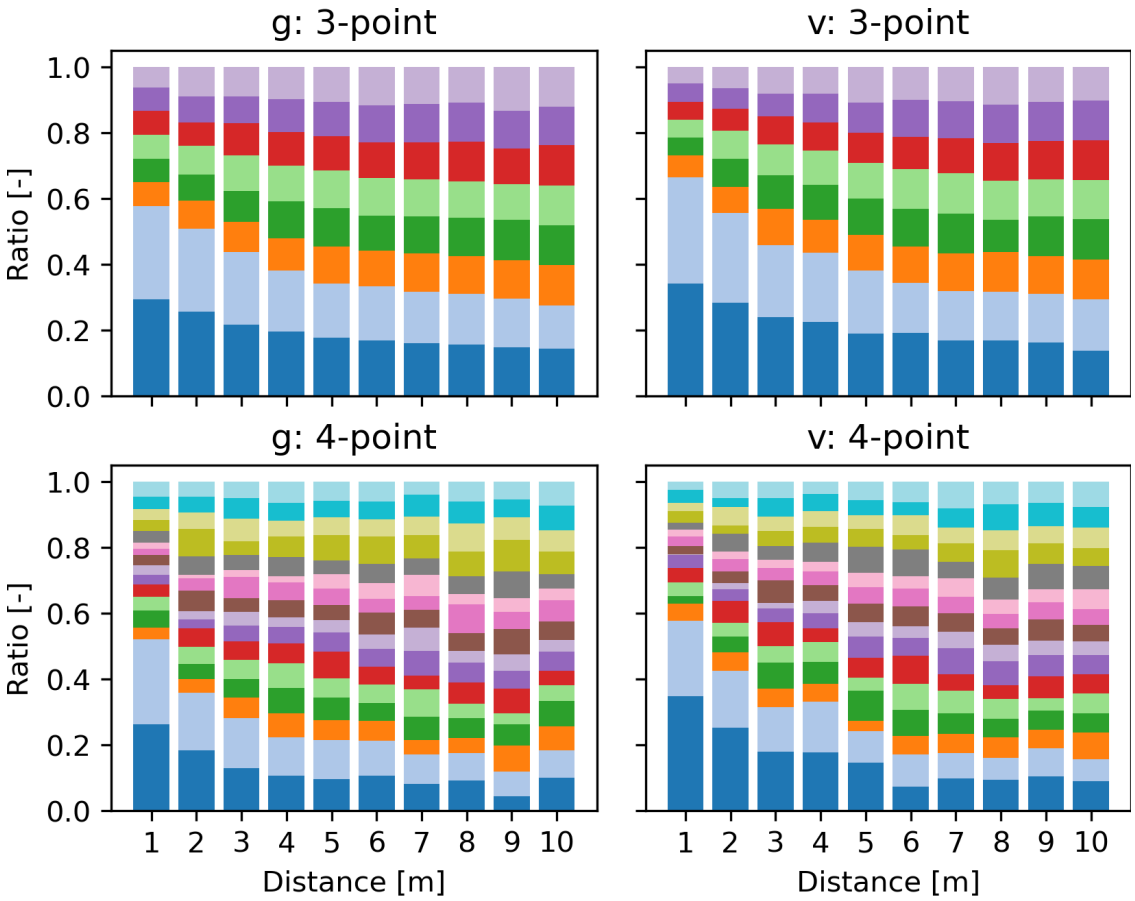

Figure 2.6: Ratio of the number of each n-point combination of 3-point (first row) and 4-point correlations (second row) of the random fields in Figure 2.4A (left column) and B (right column). Each color presents the ratio of one possible 3- or 4-point combination.

2.3.4 Variability

As discussed in the above sections, the variability of a random field is influenced by various control parameters, e.g., the variogram of the marginal distribution, the kind of multivariate distribution, the degree of non-Gaussianity, and important local features. In practice, the variability of a K-field and its influence on solute transport is controlled by several of these parameters. Therefore, the "variability" is defined in this thesis as a K-field with long-tail marginal distribution, non-multi-Gaussian underlying spatial dependence structure, and important local features. To model such K fields, copula-based geostatistical methods (see Section 2.4) are used to quantify the deviation from Gaussianity of a random field. In Chapter 3, a copula-based conditional simulation method is introduced that includes various types of information from K measurements in one simulation to simulate

realizations of K-field.

2.4 Copula Based Geostatistics

A copula C is a multivariate distribution function on the n-dimensional unit cube with uniform marginal distributions (Joe, 1997). It is inherent to any n-dimensional joint distribution function $F_{Z_1,...,Z_n}$ by replacing the values of the individual variables with their marginal cumulative probabilities $F_{Z_i}(z_i)$

$$\begin{aligned} C : [0,1]^n &\longrightarrow [0,1] \\ F_{Z_1,...,Z_n}(z_1,...,z_n) &= C(F_{Z_1}(z_1),...,F_{Z_n}(z_n)) \\ &= C(u_1,...,u_n), \end{aligned} \tag{2.17}$$

in which $u_i = F_{Z_i}(z_i)$ is the value in the copula space. Then the density function of a copula $c(u_1,...,u_n)$ can be calculated from the copula C

$$c(u_1,...,u_n) = \frac{\partial C(u_1,...,u_n)}{\partial u_1 ... \partial u_n}. \tag{2.18}$$

A copula describes the multivariate dependence structure independent of the margins (Equation 2.17). It is invariant to monotonic transformations of the marginal distribution (Sklar's theorem). Therefore, a copula is not influenced by all transformations discussed in Section 2.3.1, including the log-transformation.

The spatial dependence structure of random space functions can be evaluated using a spatial copula C_s

$$C_s(u_1,...u_n) = C(F_z(z(\mathbf{x}_1)),...,F_z(z(\mathbf{x}_n))), \tag{2.19}$$

which describe the n-point dependence structure in space as the variogram in Equation 2.9. A bivariate spatial copula can be written as (Bárdossy, 2006):

$$C_s(\mathbf{h},u_1,u_2) = P[F_z(z(\mathbf{x})) < u_1, F_z(z(\mathbf{x}+\mathbf{h})) < u_2] = C(F_z(z(\mathbf{x})), F_z(z(\mathbf{x}+\mathbf{h}))), \tag{2.20}$$

in which $\mathbf{h}$ is the separation vector between two points, $u_i = F_Z(z(\mathbf{x}_i))$ is the value of the cumulative distribution function $F_z(z(\mathbf{x}_i))$ of the data value $z(\mathbf{x})$ at location $\mathbf{x}$ in copula space.

Details of copula theory can be found in (Joe, 1997; Nelsen, 2000; Joe, 2014) and will be not be repeated in this thesis. In Section 2.4.1, the empirical bivariate copula measures are presented, which are used to describe the bivariate spatial dependence in this thesis. Theoretical spatial copula models for the high-dimensional spatial dependence

are introduced in Section 2.4.2. In Section 2.4.3, the copula parameter estimation with a censoring threshold is presented.

2.4.1 Bivariate Measures of the Spatial Dependence Based on the Copula

Data-based descriptive summary measures of the copulas can be evaluated to compare different measures. Examples of such measures are the empirical bivariate copula density c_s, the copula-based rank correlation ρ_s, and the copula asymmetry A, which indicates the degree of the deviation from symmetric Gaussian dependence, all of which evaluated for different lag distances. These empirical measures are based on data and are used to explore certain characteristics of two-point spatial dependence structures in rank space, which means that the influences of extreme values are removed. The kind of dependence can be inferred and the similarity of the spatial dependencies between datasets can be quantified using these measures.

Empirical Bivariate Copula Density

The empirical bivariate copula density c_s of n observations $z(\mathbf{x_1}),\ldots,z(\mathbf{x_n})$, with coordinate vectors $\mathbf{x_1},\ldots,\mathbf{x_n}$ is defined as the bivariate probability density of the set of data pairs $S(\mathbf{h})$ with separation vector (lag distance) $\mathbf{h}$ (Bárdossy, 2006):

$$S(\mathbf{h}) = \{(F_z(z(\mathbf{x}_i)), F_z(z(\mathbf{x}_i)))|\mathbf{x}_i - \mathbf{x}_j \approx \mathbf{h} \text{ or } \mathbf{x}_j - \mathbf{x}_i \approx \mathbf{h}\}. \tag{2.21}$$

As with the variogram computation, a lag tolerance can be introduced and data pairs with separation distances falling within the same bin (nominal lag plus or minus the tolerance) are used to represent that nominal lag. Equation 2.21 can also be applied in the anisotropic case when the spatial structure varies significantly with direction. In this case, the data pairs are grouped according to both the magnitude and the direction of $\mathbf{h}$.

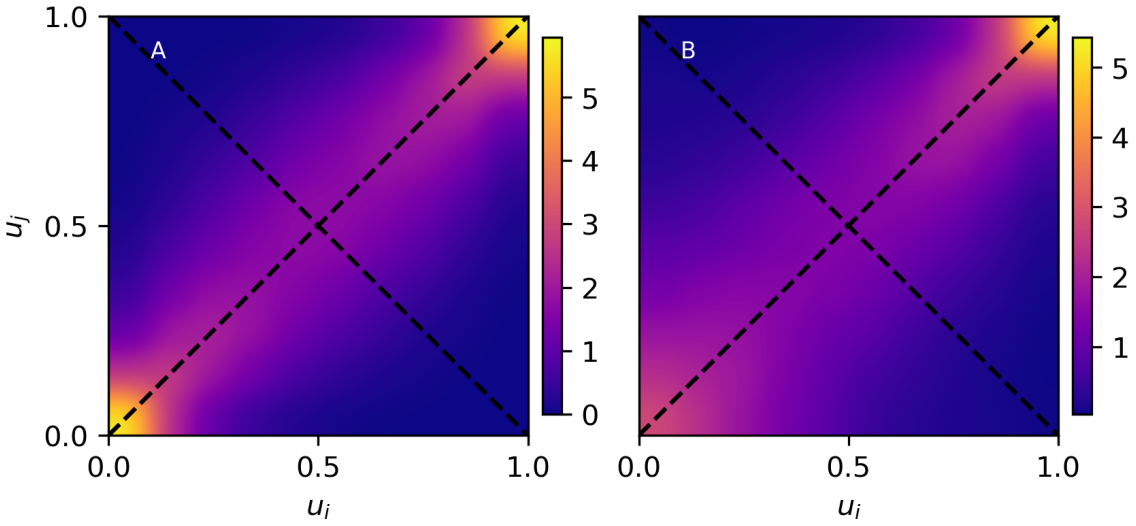

Figure 2.7: Pseudocolor plots of the empirical copula density of A) a multi-Gaussian and B) a non-multi-Gaussian random field.

A copula density c_s can be visualized using a color plot (Figure 2.7) to evaluate the information of dependence structures provided by the distribution of the data pairs $u_i = u(\mathbf{x_i}) = F_z(z(\mathbf{x_i}))$ and $u_j = u(\mathbf{x_j}) = F_z(z(\mathbf{x_j}))$ for a certain lag distance $\mathbf{h}$ and has the following properties:

1. The data pair (u_i, u_j) and the data pair (u_j, u_i) belong to a same $S(\mathbf{h})$. So, it is symmetric about $u_i = u_j$ in a non-directional spatial dependence structure.

2. c_s presents more information of data pairs regarding the relationships among data pairs than a summary measure such as the variogram, which is an average measure of variability between all data pairs with a separation distance of approximately $\mathbf{h}$. When $\mu(F_Z) = 0.5$ is defined as a threshold of large and small values. Besides the data pairs with both $u_i = 0.5$ and $u_j = 0.5$, $S(\mathbf{h})$ can be classified in four groups $\hat{S}(\mathbf{h})_i$ according to their values in the copula space:

$$\hat{S}(\mathbf{h})_1 : u_i < 0.5 \,\&\, u_j < 0.5 \tag{2.22}$$
$$\hat{S}(\mathbf{h})_2 : u_i > 0.5 \,\&\, u_j > 0.5 \tag{2.23}$$
$$\hat{S}(\mathbf{h})_3 : u_i < 0.5 \,\&\, u_j > 0.5 \tag{2.24}$$
$$\hat{S}(\mathbf{h})_4 : u_i > 0.5 \,\&\, u_j < 0.5. \tag{2.25}$$

   $\hat{S}(\mathbf{h})_1$ are plotted near the origin $(0,0)$ and $\hat{S}(\mathbf{h})_2$ are plotted near the upper right corner $(1,1)$ of the unit square. $\hat{S}(\mathbf{h})_3$ and $\hat{S}(\mathbf{h})_4$ are plotted near the upper left corner and the lower right corners.

3. In the case of no dependence, c_s has a flat structure. If there is a strong dependence structure. Areas with a high density and areas with a low density can be found in the plot.

4. A multi-Gaussian dependence (Figure 2.7A) has the same proportion of $\hat{S}(\mathbf{h})_1$ and $\hat{S}(\mathbf{h})_2$. In contrast, a non-multi-Gaussian dependence (Figure 2.7B) has asymmetric densities on the lower left and on the upper right side.

5. A sequence of copula density plots for different lags provides a representation of the spatial dependence structure along the separation distance.

More details of the bivariate copula density can be found in Guthke (2013).

Rank Correlation

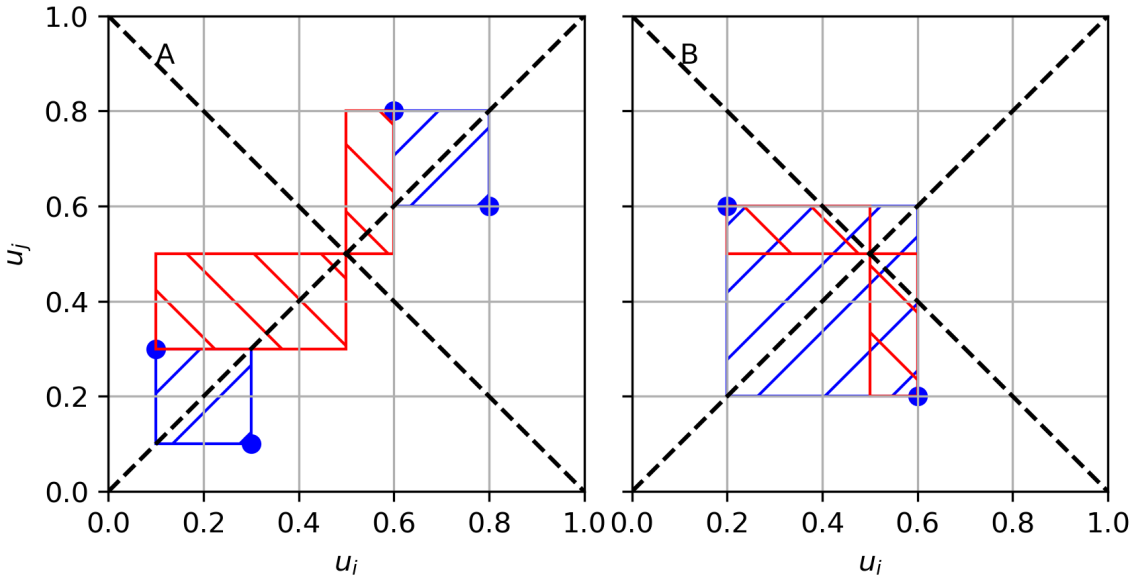

Figure 2.8: Contribution of variogram γ (blue square) and copula rank correlation ρ_s (red rectangle) of data pairs (blue dots) A) $\hat{S}(\mathbf{h})_1$ and $\hat{S}(\mathbf{h})_2$ B) $\hat{S}(\mathbf{h})_3$ and $\hat{S}(\mathbf{h})_4$.

The copula-based second-order central moment rank correlation ρ_s (Haslauer *et al.*, 2012)

$$\rho_s(\mathbf{h}) = \frac{12}{N(\mathbf{h})} \sum_{\mathbf{x_i}-\mathbf{x_j}\approx\mathbf{h}} \left(u(\mathbf{x_i})-0.5\right)\left(u(\mathbf{x_j})-0.5\right) \tag{2.26}$$

can summarize the empirical copula density c_s quantitatively, in which $N(\mathbf{h})$ is the number of data pairs for the lag distance $\mathbf{h}$.

The behavior of ρ_s is similar to the behavior of a normalized covariogram $Cov(\mathbf{h}) = 1 - \frac{\gamma(\mathbf{h})}{\sigma^2}$. For very short lag distances, ρ_s is unity, and it decreases until it reaches zero, indicating a lack of correlation between data pairs separated by a sufficiently large distance. However, since ρ_s is in the copula space, it is less influenced by extreme values. Figure 2.8 shows the corresponding values of the variogram (blue square; Equation 2.9) and the rank correlation (red rectangle) of different types of data pairs in the copula space. A data pair always contributes a positive term in the variogram, which only depends on the difference of these two points. Data pairs $\hat{S}(\mathbf{h})_1$ and $\hat{S}(\mathbf{h})_2$ (Figure 2.8A) contribute a positive term in ρ_s and data pairs $\hat{S}(\mathbf{h})_3$ and $\hat{S}(\mathbf{h})_4$ (Figure 2.8B) contribute a negative term in ρ_s. A data pair away from the center point $(0.5, 0.5)$ has a larger positive or smaller negative weight than a data pair close to the middle point (Figure 2.8).

The rank correlations of some well-known K datasets have been calculated, including those from Borden, Canada (Sudicky, 1986), Cape Cod, USA (LeBlanc *et al.*, 1991), North Bay (NB), Canada (Sudicky *et al.*, 2010), and the MADE site, USA (Boggs *et al.*, 1992; Bohling *et al.*, 2016). All these datasets exhibit similar second-order dependence with different ranges (Figure 2.9)

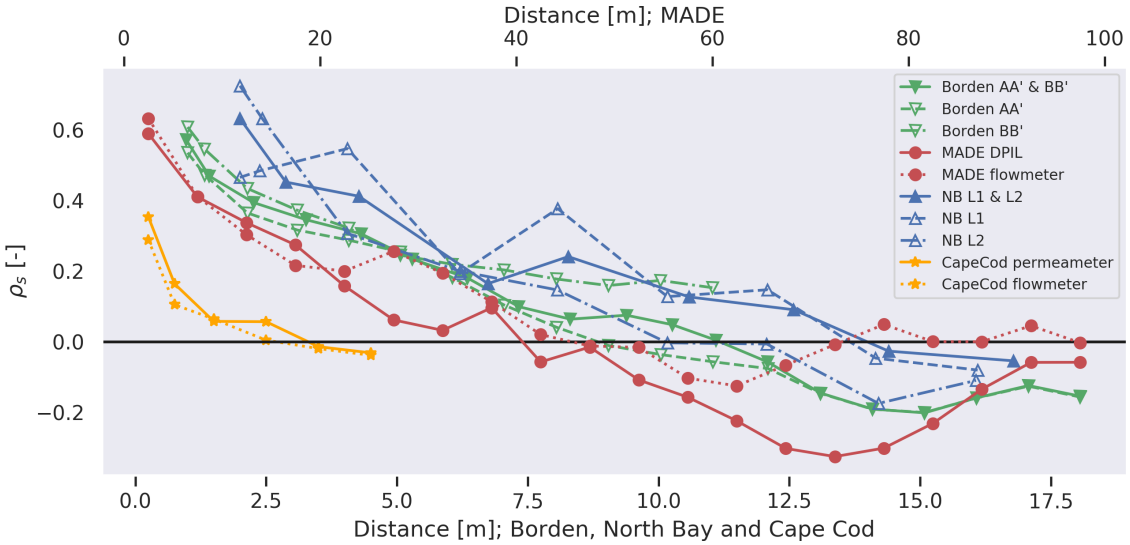

Figure 2.9: Correlograms versus separation distance for common datasets in stochastic hydrogeology.

Asymmetry

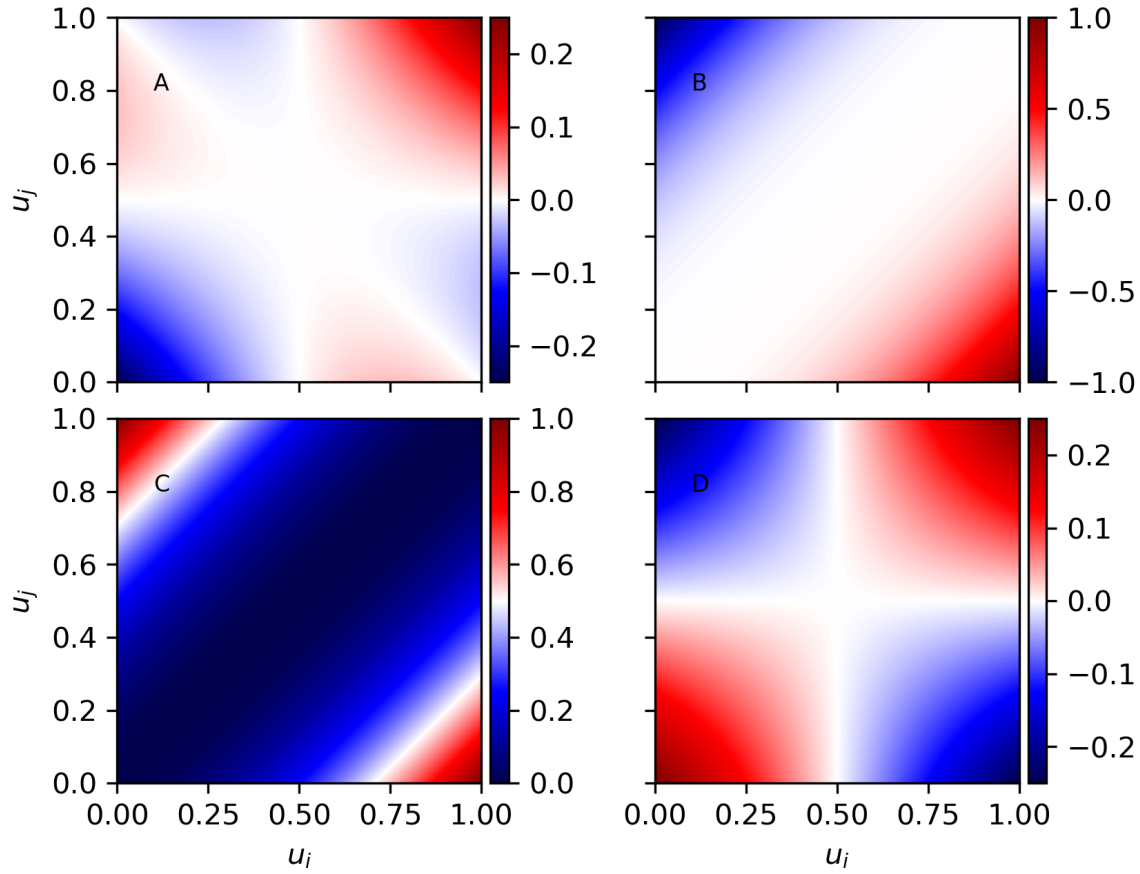

Figure 2.10: Pseudocolor plots of the contributions of different data pairs in A) Asymmetry in Equation 2.28; B) Asymmetry in Equation 2.29 C) variogram in Equation 2.9, and D) rank correlation in Equation 2.26

The copula density of a multi-Gaussian spatial dependence structure is symmetric for all separation distances. So the "degree of Gaussianity" of a spatially distributed variable can be quantified by the deviation of the copula density of a dataset from a symmetric (standard normal) copula density. This measure is defined as the third-order spatial measure asymmetry (A) in copula space, which is either based on data or a theoretical model (Haslauer *et al.*, 2012):

$$A(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{\mathbf{x_i}-\mathbf{x_j}\approx\mathbf{h}} \hat{A}(u(\mathbf{x_i}), u(\mathbf{x_j})), \tag{2.27}$$

in which $N(\mathbf{h})$ is the number of data pairs within a certain range centered about the separation vector $\mathbf{h}$ and $\hat{A}\left(u(\mathbf{x_i}), u(\mathbf{x_j})\right)$ is the contribution of an individual data pair to

the asymmetry A. The empirical asymmetry A used in this thesis is defined as (Haslauer *et al.*, 2012):

$$\hat{A}\left(u(\mathbf{x_i}),u(\mathbf{x_j})\right) = \Big(\left(u(\mathbf{x_i})-0.5\right)^2\left(u(\mathbf{x_j})-0.5\right)\Big) + \Big(\left(u(\mathbf{x_i})-0.5\right)\left(u(\mathbf{x_j})-0.5\right)^2\Big). \tag{2.28}$$

$\hat{A}(\hat{S}(\mathbf{h})_3)+\hat{A}(\hat{S}(\mathbf{h})_4)=0$. So, the asymmetry in Equation 2.28 shows the difference between the densities of data pairs $\hat{A}(\hat{S}(\mathbf{h})_1)$ and $\hat{A}(\hat{S}(\mathbf{h})_2)$, i.e., data pairs with both small values and both large values, on the two sides of the line $u_2 = 1-u_1$, the line from the upper left corner to the lower right corner of the bivariate density plot. This property can be used to reduce the computational cost of the empirical asymmetry by calculating $\hat{A}(\hat{S}(\mathbf{h})_1)$ and $\hat{A}(\hat{S}(\mathbf{h})_2)$ and drop $\hat{A}(\hat{S}(\mathbf{h})_3)$ and $\hat{A}(\hat{S}(\mathbf{h})_4)$. As an example, for a four-point combination with the corresponding values $z=(0.1,0.2,0.8,0.9)$, the required number of data pairs to calculate is reduced from 6 to 2.

Depending on the properties of the dataset, an asymmetry can be defined in different ways. Another type of asymmetry is the directional asymmetry (Bárdossy and Hörning, 2017):

$$\hat{A}\left(u(\mathbf{x_i}),u(\mathbf{x_j})\right) = \left(u(\mathbf{x_i})-u(\mathbf{x_j})\right)^3, \tag{2.29}$$

in which $\hat{A}(\hat{S}(\mathbf{h})) = -\hat{A}(\hat{S}(-\mathbf{h}))$.

For a conceptual visualization of the represented information, Figure 2.10 shows the color plots of the values of different bivariate measures in the copula space and presents the corresponding values of different data pairs from the data. The asymmetry in Equation 2.28 (Figure 2.10A) calculate mainly the difference between $\hat{A}(\hat{S}(\mathbf{h})_1)$ and $\hat{A}(\hat{S}(\mathbf{h})_2)$ and the asymmetry in Equation 2.29 (Figure 2.10B) calculate mainly the difference between $\hat{A}(\hat{S}(\mathbf{h})_3)$ and $\hat{A}(\hat{S}(\mathbf{h})_4)$. The variogram in Equation 2.9 has nonnegative values in the whole space, in which the data pairs $\hat{A}(\hat{S}(\mathbf{h})_3)$ and $\hat{A}(\hat{S}(\mathbf{h})_4)$ close to the corners have large values. In contrast, the rank correlation in Equation 2.26 presents the difference between $\hat{A}(\hat{S}(\mathbf{h})_1)$, $\hat{A}(\hat{S}(\mathbf{h})_2)$ and $\hat{A}(\hat{S}(\mathbf{h})_3)$, $\hat{A}(\hat{S}(\mathbf{h})_4)$. Different measures extract different information from the data. The selection of the bivariate measure in practice depends on the properties of the dataset and the properties of the process relying on the dataset.

The asymmetry can be calculated for various separation distances. A positive asymmetry indicates stronger dependence among larger quantiles than smaller quantiles and negative asymmetry indicates the opposite. Depending on the correlation, a typical value of A in Equation 2.28 is between -0.03 and 0.03. As a special case, a perfectly multi-Gaussian dependence has equal dependence in both high and low quantiles and is symmetric about the line $u_2 = 1-u_1$. The value of A in this case is zero along the lag distance. Thus,

a value of A that is consistently different from zero is evidence for the existence of a non-Gaussian dependence structure.

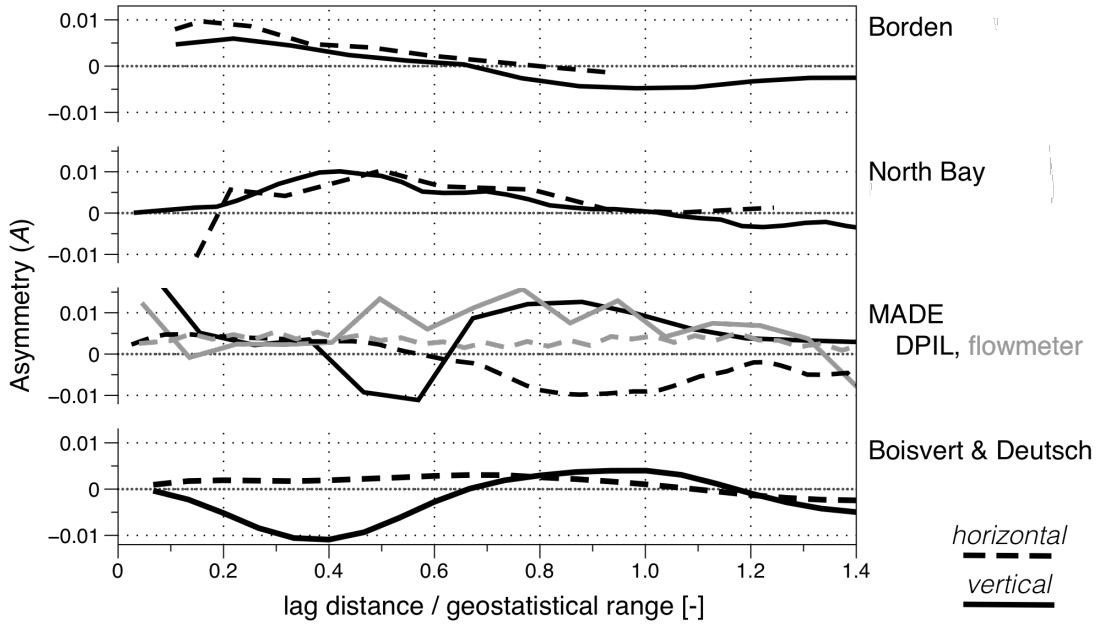

Figure 2.11: Asymmetry (A) of the hydraulic-conductivity copula as a function of the normalized lag-distance at the Borden site (Sudicky, 1986), the North-Bay site (Sudicky *et al.*, 2010), the MADE site (Boggs *et al.*, 1992; Bohling *et al.*, 2016), and in the study of Boisvert and Deutsch (2011). The lag distance is normalized by the range of the variogram.

Asymmetry is typically occurring as a consequence of diffusive processes (Guthke and Bárdossy, 2017). In practice, a spatially distributed dataset can exhibit both positive and negative asymmetry depending on the separation distance, and the zero-line may be crossed. This has been observed in many hydrogeological datasets. Figure 2.11 shows the asymmetry function of hydraulic conductivity at some of the best-studied sites worldwide. Varying values of asymmetry describe a varying dependence structure, indicated by a varying degree of dependence in varying quantiles. For example, large and small values could be connected differently, indicating different types of non-Gaussianity. For more background on copulas and their asymmetry, please refer to Bárdossy and Pegram (2009); Sugimoto *et al.* (2016); Guthke and Bárdossy (2017); Hörning and Bárdossy (2018).

2.4.2 Theoretical Spatial Copulas

This section presents a summary of theoretical spatial copula models: the Gaussian copula and the v-copula and shows how to estimate the parameters from a dataset. A spatial copula model must satisfy certain conditions beyond those required in a non-spatial multivariate context. As defined by Haslauer *et al.* (2012), these additional conditions are:

1. The bivariate spatial copula depends only on the lag vector **h** between observations and is independent of the location of the observations (assumption of stationary).

2. An n-dimensional spatial copula can be built using any n-point subset of the observations and describes the spatial dependence among these observations.

3. An arbitrarily strong dependence has to be modeled.

A spatial Gaussian copula $C_{s,\Phi}$ and the corresponding Gaussian copula density $c_{s,\Phi}$ can be constructed from the n-dimensional multivariate Gaussian distribution Φ_n with a covariance matrix $\mathbf{\Gamma}_n$ and its margins $\Phi(u_i)$ (Haslauer, 2011):

$$C_{s,\Phi}(u_1,...,u_n) = \Phi_n(\Phi^{-1}(u_1),...,\Phi^{-1}(u_n)) \tag{2.30}$$

and

$$c_{s,\Phi}(u_1,...,u_n) = \frac{1}{\sqrt{\mid \mathbf{\Gamma}_n \mid}}\Big(-\frac{1}{2}(\mathbf{\Phi}_n^{-1})^T(\mathbf{\Gamma}_n^{-1}-\mathbf{I})\mathbf{\Phi}_n^{-1}\Big). \tag{2.31}$$

Different copula models for non-multi-Gaussian spatial dependence have been developed, including the v-copula and the maximum Gauss copula (Haslauer *et al.*, 2012), among others. The v-copula model is used in this thesis, which is based on a non-monotonic v-transformation $f_v(Y,m_c,k_c)$ of a standard normal variable Y:

$$X_j = f_v(Y_j,m_c,k_c) = \begin{cases} k_c(Y_j-m_c), & if\, Y_j \geq m_c \\ m_c - Y_j, & otherwise, \end{cases} \tag{2.32}$$

in which m_c, $k_c \in \mathbb{R}^+$ are the two model parameters of the v-transformation and X is the v-transform of an n-dimensional normal random variable $\mathbf{Y} : \Phi(\mathbf{0},\mathbf{\Gamma})$ with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Gamma}$. When $m_c \gg 3$, the v-copula approximates a multi-Gaussian model. The reverse form of the v-transformation can be calculated using $\Phi^{-1}(1-F_Z(z))$ as the input value to simulate a higher correlation within small values. Figure 2.12 shows how a multi-Gaussian random field is changed by a v-transformation with $m_c = 1.0$ and $k_c = 2.0$. The order of the data in the rank space is not consistent anymore after the v-transformation (Figure 2.12B). This non-monotonic v-transformation transforms not only the marginal distribution from a symmetric distribution to a nonsymmetric

distribution but also how the random field looks like in space (Figure 2.12C-F). Before the v-transformation, large values (orange blobs) and small values (blue blobs) have a similar character in the random field (Figure 2.12E, orange and blue blobs with equal size). The large values in the v-transformed random field are more close to each other than the small values (Figure 2.12F, orange blobs are larger than blue blobs). This v-transformed field can be used to model a non-mullti-Gaussian K-field with different behavior between large and small K-values.

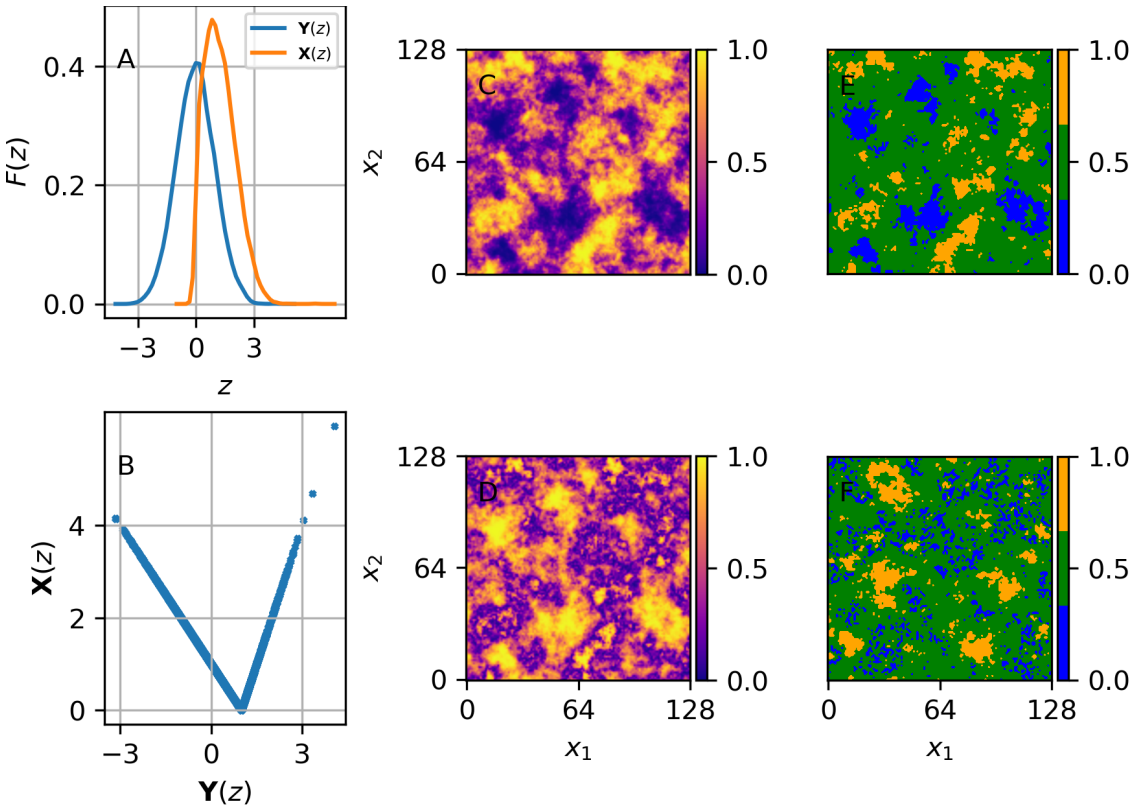

Figure 2.12: A) A normal density function before ($\mathbf{Y}(z)$) and after ($\mathbf{X}(z)$) the v-transformation with $m_c = 1.0$ and $k_c = 2.0$; B) The scatter plot between $\mathbf{Y}(z)$ and $\mathbf{X}(z)$; C) Spatial distribution of $F_z(\mathbf{Y}(z))$ and D) spatial distribution of $F_z(\mathbf{X}(z))$; E) Indicator plot of $F_z(\mathbf{Y}(z))$ and F) $F_z(\mathbf{X}(z))$.

The v-transformed marginal distribution function $F_Z(x)$ and marginal density function $f_Z(x)$ can be calculated using (Haslauer, 2011):

$$F_Z(x) = \Phi\left(\frac{x}{k_c} + m_c\right) - \Phi(-x - m_c) \tag{2.33}$$

and

$$f_Z(x) = \frac{1}{k_c}\phi(\frac{x}{k_c}) + \phi(-x - m_c). \tag{2.34}$$

The v-transformed multivariate distribution function F_n and f_n can be calculated using (Haslauer, 2011)

$$F_n(x_1, ...x_n) = \sum_{i=0}^{2^{n-1}} (-1)^{n-\sum_{j=0}^{n-1} i_j} \mathbf{\Phi}(\zeta_i + \mathbf{m}) \tag{2.35}$$

and

$$f_n(x_1, ...x_n) = \frac{1}{\sqrt{(2\pi)^n \mid \Gamma_n \mid}} \sum_{i=0}^{2^n-1} \big[\frac{1}{k^{n-\sum_{j=0}^{n-1} i_j}} \cdot \exp\big(-\frac{1}{2}(\zeta_i + \mathbf{m})^T \mathbf{\Gamma}_n^{-1}(\zeta_i + \mathbf{m})\big)\big], \tag{2.36}$$

in which

$$\zeta_i^T = \big(b((-1)^{i_1}) \cdot x_1, ..., b((-1)^{i_n}) \cdot x_n\big), \tag{2.37}$$

$$i = \sum_{j=0}^{n-1} i_j 2^j, \tag{2.38}$$

$$b = \begin{cases} -1 & if (-1)^{i_j} = -1 \\ \frac{1}{k} & if (-1)^{i_j} = +1. \end{cases} \tag{2.39}$$

Then the v-copula density $c_{s,v}$ can be calculated from the joint multivariate density function $f_n(x)$ and the marginal density function $f(x)$ using (Haslauer, 2011):

$$c_{s,v}(u_1, ...u_n) = \frac{f_n(x_1, ...x_n)}{\prod_{i=1}^{n} f_i(x_i)}. \tag{2.40}$$

The maximum likelihood (ML) based approach described in Bárdossy and Li (2008) is used in this thesis to estimate the Gauss- and v-copula model parameters. The dataset S is separated into w small subsets S_w, each with $n(S_w)$ data points and $n(S_w) \gg 2$, because of the complexity of the computation. Then the correlation matrix $\mathbf{\Gamma}_{S_w}$ for each subset is calculated as

$$\mathbf{\Gamma}_{S_w} = \left(\left(R_{S_{w_i}, S_{w_j}}(\mathbf{h})\right)^{n(S_w), n(S_w)}\right), \tag{2.41}$$

in which $R_{S_{w_i}, S_{w_j}}$ is the correlation function, which depends only on the vector $\mathbf{h} = \mathbf{x}_{S_{w_i}} - \mathbf{x}_{S_{w_j}}$. This correlation function can be calculated by superposition of one or more correlation functions as

$$R_{S_{w_i}, S_{w_j}}(\mathbf{h}) = \sum_{k=0}^{K} D_k r_k(\mathbf{h}, a_k) \text{ with } \sum_{k=0}^{K} D_k = 1, \tag{2.42}$$

in which $r_k(\mathbf{h}, a_k)$ is the correlation function related to the k-th theoretical variogram, D_k is the weight of the correlation function and a_k is the range of the correlation function. The likelihood for each subset can be calculated from the corresponding copula density with parameter vector $\boldsymbol{\theta}$ (Equation 2.43)

$$c_s(S_w, \boldsymbol{\theta}) = c_s(F_n(Z(x_{S_{w_1}})), ..., F_n(Z(x_{S_{w_{n(S_w)}}})), \boldsymbol{\theta}) \tag{2.43}$$

and

$$L(\boldsymbol{\theta}|Z(x_1), ..., Z(x_n)) = \prod_{w=1}^{W} c(S_w, \boldsymbol{\theta}) \rightarrow max. \tag{2.44}$$

The likelihood $L(\boldsymbol{\theta}|Z(x_1), ..., Z(x_n))$ in Equation 2.44 is maximized by varying $\boldsymbol{\theta}$. This process is repeated several times to reduce the influence of the choice of the subset. Finally, the average value of all estimation results is calculated, yielding the estimated copula parameters m_c and k_c.

2.4.3 Theoretical Spatial Gaussian Copula with Censored Measurements

To take censored data into account, censoring thresholds between 0 and 100 percentage in CDF space are defined as input parameters. The probability space is split into two parts for a one-side censoring threshold or three parts for a two-side censoring threshold. If the threshold represents an upper limit, data values below the threshold are considered crisp and those above it to be censored, and vice versa if it represents a lower limit. Then a two-step maximum likelihood parameter estimation is performed.

The first step is the calculation of the cumulative distribution function using the kernel density estimation method. While calculating the likelihood function, the values of the probability density function for the censored data are replaced by the area of the part below the lower threshold $F_Z(DLL) - F_Z(0)$ or the area of the part above the upper threshold $1 - F_Z(DLR)$ (Haslauer *et al.*, 2017b).

The second step is the copula parameter estimation. The values of the censored data are replaced by the conditional copula density based on the crisp data (Bárdossy, 2011). Then, the likelihood function can be written as

$$L(\boldsymbol{\theta}|Z(x_1), ..., Z(x_n)) = \left(\prod_{w_1=1}^{W_1} c_s(S_{w_1}, \boldsymbol{\theta}) \right) \cdot \left(\prod_{w_2=1}^{W_2} c_s(S_{w_2}, \boldsymbol{\theta}) \right), \tag{2.45}$$

in which w_1 is the subset with only crisp data, and w_2 is the subset with both crisp and

censored data.

Chapter 3

Copula-Based Geostatistical Conditional Simulation

The goal of the copula-based geostatistical simulation methodology in this thesis is to incorporate different types of univariate point measures, and second-order two-point spatial measures, namely the variogram and the asymmetry of the two-point copula, all derived from the available K observations, in a unifying stochastic model. A key point to achieve this goal is to include the measure of Gaussianity in a three-dimensional simulation by an efficient computational algorithm. When a copula-based parameter interface is available, (Gaussian copula, v-copula (Chapter 2.4) or maximum Gauss copula (Haslauer *et al.*, 2012)) parameter estimation can be performed on the observations to get the required parameters for the simulation. Bárdossy and Hörning (2016) include the v-copula in the random mixing algorithm by using the v-transformation (Equation 2.32) for inverse modeling of groundwater flow. Li (2010) used the sequential Gaussian simulation to calculate the conditional copula with the corresponding parameterized copula model for a conditional simulation. One disadvantage of this group of methods is that there is a limited number of theoretical copula models that cannot describe all possible kinds of non-multi-Gaussian dependence structures that might exist in the real world. Data-based empirical measures can describe certain properties in data with the disadvantage that they are not a full model and are computationally more intensive. Hörning and Bárdossy (2018) use the phase-annealing (PA=**P**hase randomization + Simulated **A**nnealing) algorithm to simulate realizations that exhibit the prescribed empirical measures (asymmetry in Section 2.4.1). Lauzon and Marcotte (2019) include the asymmetry in the Fast Fourier Transform Moving Average-simulated annealing (FFTMA-SA) for the calibration of a random field. The theoretical background of the phase-annealing method is introduced in this chapter and it is applied to the K-field simulation at the MADE site in Chapter 6. A brief introduction of the Simulated Annealing (SA) method is presented in Section 3.1. Then PA is introduced in Section 3.2. Section 3.3 and Section 3.4 show how to include the Gaussianity in phase-annealing with v-transformation and the FFT-based method. At the end of this chapter, some related computational aspects of PA are discussed in Section 3.5.

3.1 Simulated-Annealing with Asymmetry

The simulated-annealing method was first used in statistical mechanics to model the interaction of a many-body system. Later it has been widely used as a tool for global stochastic optimization of large combinatorial problems (Kirkpatrick *et al.*, 1983) and extended to model a spatial image (Geman and Geman, 1984).

A SA implementation for the simulation of a spatially distributed random field typically consists of the following steps:

1. setup an initial status of the random field as the start point of the annealing iterations,
2. a perturbation mechanism to update the current status (RF perturbation),
3. an objective function (obj) to calculate the distance between the current status and the required status,
4. acceptance or rejection of the current perturbation.

SA is flexible to solve various problems and can include a variety of objective functions. One disadvantage of SA is the high computational cost when many iterations are needed or the perturbation/updating mechanism of the objective function requires large computational resources. In geostatistical simulations, the updating of the variogram and asymmetry cost much computational time, especially for a three-dimensional simulation. Deutsch and Cockerham (1994) showed a method that reduces the computational cost of variogram updating by swapping pairs of point values as the perturbation mechanism. The variogram $\gamma_{new}(\mathbf{h})$ after the point swapping is not calculated fully in each step 3 but updated from $\gamma_{old}(\mathbf{h})$ by treating the variogram contributions of the swapped points (Deutsch and Cockerham, 1994):

$$\gamma_{new}(\mathbf{h}) = \gamma_{old}(\mathbf{h}) - [z(\mathbf{x}+\mathbf{h}) - z(\mathbf{x})]^2 + [z(\mathbf{x}+\mathbf{h}) - z'(\mathbf{x})]^2, \tag{3.1}$$

in which $z(\mathbf{x})$ is the value before swapping and $z'(\mathbf{x})$ is the value after the swapping. In the case of asymmetry, the distribution function $F_Z(\mathbf{x})$ is constant during the point swapping. So the perturbation mechanism can be extended to update the asymmetry A_{new} from the

A_{old}:

$$\begin{aligned} A_{new}(\mathbf{h}) = A_{old}(\mathbf{h}) & \\ & - \left(\left(F_z(\mathbf{x}) - \frac{1}{2} \right)^2 \left(F_z(\mathbf{x}+\mathbf{h}) - \frac{1}{2} \right) \right) \\ & - \left(\left(F_z(\mathbf{x}) - \frac{1}{2} \right) \left(F_z(\mathbf{x}+\mathbf{h}) - \frac{1}{2} \right)^2 \right) \\ & + \left(\left(F_z'(\mathbf{x}) - \frac{1}{2} \right)^2 \left(F_z(\mathbf{x}+\mathbf{h}) - \frac{1}{2} \right) \right) \\ & + \left(\left(F_z'(\mathbf{x}) - \frac{1}{2} \right) \left(F_z(\mathbf{x}+\mathbf{h}) - \frac{1}{2} \right)^2 \right), \end{aligned} \tag{3.2}$$

in which $F_z(\mathbf{x})$ and $F_z'(\mathbf{x})$ are the distribution function values before and after the swapping. A non-multi-Gaussian random field can be simulated with asymmetry using SA as in Algorithm 1.

3.2 Phase-Annealing

Phase-annealing (Hörning and Bárdossy, 2018) is a modified variant of SA and uses phase randomization as the perturbation mechanism.

3.2.1 Phase Randomization

The advantage of the SA method is its flexibility to combine different types of the objective function and disparate sources of data in one optimization process (Deutsch and Cockerham, 1994). The direct observations of the random field at observation points, in SA are set as fixed points during a conditional simulation to obtain realizations matching the observations. This could lead to a non-harmonic structure between the conditional point and its neighboring points. Hörning and Bárdossy (2018) discussed the effect of these singularities and suggested replacing the original point-swap perturbation mechanism with the phase randomization (PR) method in the Fourier space.

The Wiener-Khinchin theorem connects the covariance function *Cov* of a random process

Algorithm 1: Simulated-annealing with asymmetry.

Result: RF_{out} with the required properties
initialization;
the initial status of the RF is a white noise random field $\Phi(0,1)$;
while $N_1 < N_{max,iter1}$ *and* $obj > obj_{tol}$ **do**
  $T \leftarrow T_0 \cdot T_{frac}^{N_1}$ update annealing temperature with the cooling ratio T_{frac} ;
  while $N_2 < N_{max,iter2}$ *and* $obj > obj_{tol}$ **do**
    RF perturbation by point swapping;
    $\gamma(\mathbf{h})_{new}(Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})) \leftarrow$ update variogram ;
    $A(\mathbf{h})_{new}(F_Z(\mathbf{x}), F_Z(\mathbf{x}+\mathbf{h})) \leftarrow$ update asymmetry ;
    $obj_{new} \leftarrow$ update objective function;
    if $obj_{new} < obj_{best}$ **then**
      $obj, obj_{best} \leftarrow obj_{new}$;
    else if $random(0,1) \leq P_{accept}(obj, obj_{new}, T)$ **then**
      $obj \leftarrow obj_{new}$;
    else
      reject current perturbation;
    end
    $N_2 \leftarrow N_2 + 1$;
  end
  $N_1 \leftarrow N_1 + 1$;
end
$RF_{out} \leftarrow$ transform to the required F_Z ;

and its power-spectral density S in the Fourier space:

$$Cov_{XX}(\tau) = \int_{-\infty}^{\infty} S_{XX}(f) e^{i2\pi f\tau} df. \tag{3.3}$$

The discrete Fourier Transform (**TF**) and its inverse ($\mathbf{TF}^{-1}$) of a one-dimensional spatial variable $\mathbf{Y}$ with N values are defined as:

$$F_{\mathbf{Y},k} = \mathbf{TF}(\mathbf{Y}) = \sum_{n=0}^{N-1} y_n \exp\left(-\frac{2\pi i}{N} kn\right), \tag{3.4}$$

$$y_n = \mathbf{TF}^{-1}(F_{\mathbf{Y}}) = \frac{1}{N} \sum_{k=0}^{N-1} F_{\mathbf{Y},k} \exp\left(\frac{2\pi i}{N} kn\right), \tag{3.5}$$

in which $F_{\mathbf{Y},k}$ is the k-th term of the Fourier coefficient in the Fourier space, y_n is the n-th term of $\mathbf{Y}$ and i is the imaginary unit. The power-spectral density $S_{\mathbf{Y},k}$ and phase angle ψ_k of one Fourier coefficient $F_{\mathbf{Y},k}$ can be defined as:

$$S_k = \frac{1}{N^2} F_{\mathbf{Y},k} \bar{F}_{\mathbf{Y},k} \tag{3.6}$$

$$\psi_k = \arctan(\mathrm{Im}(F_{\mathbf{Y},k}), \mathrm{Re}(F_{\mathbf{Y},k})) \tag{3.7}$$

in which $\bar{(\cdot)}$, $\mathrm{Im}(\cdot)$, and $\mathrm{Re}(\cdot)$ are the complex conjugate, imaginary component, and real component of a complex argument, respectively; S_k is the spectral power (or squared amplitude), and ψ_k is the phase angle. The spectral power S_k after shifting a phase angle ψ_k to ψ_k^* is identical to the spectral power before shifting the phase angle:

$$\begin{aligned}(S_k^*)^2 =& \mathrm{Re}\left(S_k \cdot \exp(-i \cdot \psi_k^*)\right)^2 + \mathrm{Im}\left(S_k \cdot \exp(-i \cdot \psi_k^*)\right)^2 \\ =& \left(S_k \cdot \cos(-\psi_k^*)\right)^2 + \left(S_k \cdot \sin(-\psi_k^*)\right)^2 \\ =& S_k^2.\end{aligned} \tag{3.8}$$

The contribution of a spectral power S_k in the spatial domain is:

$$\Delta Y(\psi) = 2 \cdot \frac{1}{N} \mathrm{Re}\left(S_k(\psi) \cdot \exp\left(\frac{2\pi i}{N} kn\right)\right). \tag{3.9}$$

The updated value Y^* can be calculated as (Hörning and Bárdossy, 2018):

$$Y^* = Y - \Delta Y(\psi) + \Delta Y(\psi^*), \tag{3.10}$$

in which $\Delta Y(\psi)$ is the current contribution of the selected phase and $\Delta Y(\psi^*)$ is the contribution of the selected phase after randomly shifting the phase angle. When the phase angle ψ_k is shifted uniformly between $[0, 2\pi]$, Y can be updated by keeping S_k as a

constant, which means that the two-point covariance function is not changed. This method is called phase randomization (PR) and can be extended to n dimensions.

3.2.2 Objective Function in Phase-Annealing

Phase randomization brings two changes in the simulation process:

1. The power spectrum of a random field remains untouched when the phase spectrum is updated. So, the two-point covariance is not changed during the annealing iterations (Equation 3.8). When a random field with the required variogram is used as the initial status, the variogram term can be removed from the objective function. In this thesis, a multi-Gaussian random field is simulated first as the input field of PA because it is computationally efficient, but this is not compulsive.

2. PR is a global update method, which means the values of the entire domain are changed during each annealing iteration. So the conditional points can not be set as fixed points and need to be included in the objective function. And the asymmetry can not be updated directly using Equation 3.2.

Different types of information can be included in the simulation simultaneously: the measurement values at a set of measurement locations ("point or pixel values"), the order within a set of measurements at another non-colocated measurement locations ("orders of the point or pixel values"), and some measures for the spatial dependence structure, that might be different from second-order measures, e.g., asymmetry A as metric of non-multi-Gaussianity. The simulation problem can be treated as a multi-objective optimization. The goal of multi-objective optimization is to determine the Pareto-Front of multiple objective functions. The details of this part are not discussed in this thesis.

The required measures are combined in a single final objective function obj by a weighted-sum approach (Jahn *et al.*, 1992):

$$obj = \sum_{n=1}^{N} \omega_n obj_n, \tag{3.11}$$

in which N is the number of criteria included in the overall objective function and obj_n is the n-th individual objective function with weight ω_n. The selection of the weight ω_n depends on the problem itself. In this thesis, equalized weights $\omega_1 = \omega_2 = ... = \omega_n$ are used.

As explained above, the objective function is minimized by the phase-annealing method and consists of a weighted sum of individual objective functions, which are:

1. the point-value criterion obj_1:

$$obj_1 = \frac{MSE(y_c(\mathbf{x_i}), y^*(\mathbf{x_i}))}{obj_{1,init}}, \tag{3.12}$$

in which $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^*)^2$ is the mean squared error operator, y_c is the measured value at a point and $obj_{1,init}$ is the initial value of this type of objective function.

2. the point-order-value criterion obj_2:

$$obj_2 = \frac{\sum_{ij}\mathbf{M}(y^*(\mathbf{x_i}), y^*(\mathbf{x_j})) \cdot \omega_2(\mathbf{x_i} - \mathbf{x_j})}{obj_{2,init}}, \tag{3.13}$$

in which $\mathbf{M}$ is a logical matrix with:

$$\mathbf{M_{ij}} = \begin{cases} 0 & \text{if order is correct} \\ 1 & \text{if order is incorrect,} \end{cases} \tag{3.14}$$

in which $\omega_2(\mathbf{x_i} - \mathbf{x_j})$ is a weighting function based on the distance of two values and $obj_{2,init}$ is the initial value of this type of objective function. Furthermore, a detection limit can be integrated into the optimization by defining a second type of objective function to present the rank behavior between the simulated values and the detection limit.

3. the asymmetry-related criterion obj_3:

$$obj_3 = \frac{MSE(A_c(\mathbf{h}), A^*(\mathbf{h}), \omega_3(\mathbf{h}))}{obj_{3,init}}, \tag{3.15}$$

in which $A_c(\mathbf{h})$ is the target asymmetry, $A^*(\mathbf{h})$ is the simulated asymmetry, $\omega_3(\mathbf{h})$ is a weight function by lag distances and $obj_{3,init}$ is the initial value of this type of objective function.

Besides the asymmetry A, other types of spatial measures could be included in the optimization, e.g., cross-covariance, variogram on a different scale, measures of the connectivity, etc.

Depending on the knowledge about the given problem, e.g., the field site and the required properties of the simulated fields, the possible types of the objective function are not limited to those listed above; for example, the mean values and variances of certain locations can be used as a part of the objective function.

The objection functions in PA quantify the likelihood between the simulated values and the prior understanding of the system. Therefore it is possible to include the uncertainty in the MC simulation using a fuzzy membership function for a certain type of objective function (Abebe *et al.*, 2000).

3.2.3 Point Values and Order of Point Values as Conditional Points

A straightforward method to merge two datasets of the same parameter that have a different distribution function is the QQ-transformation (Equation 2.15), in which two datasets are connected by the CDF. The QQ-transformation assumes that both datasets sample the whole distribution function of the same random variable, which is very hard to be fulfilled for a large-scale heterogeneous space. Otherwise, the transformed values are shifted in the CDF space and a deviation is included in the system by the QQ-transformation.

As an example, Figure 3.1A shows the PDFs of

1. z_1: as a reference dataset with $z_1 = \Phi(0,1)$;
2. z_2: a subset of z_1, in which the values between $(1,2]$ are not sampled;
3. z_3: observations of z_2 with $z_3 = 2z_2 + 2$.

A small peak can be found on the right side of $f_Z(z_2)$ and $f_Z(z_3)$ because of the unsampled area in z_2.

A QQ-transformation can be performed on z_3 to transform z_3 to z_1. Because z_3 is generated from z_2. Therefore, the transformed z_3 should be the same as z_2. However, a deviation can be found in the scatter plot of z_2 and the transformed z_3 (Figure 3.1B). To reduce this deviation, which is included by using direct observations, i.e., the point values, the order of the point values can be used as additional information in PA. Because z_2 and the transformed z_3 are consistent in the rank space (Figure 3.1C).

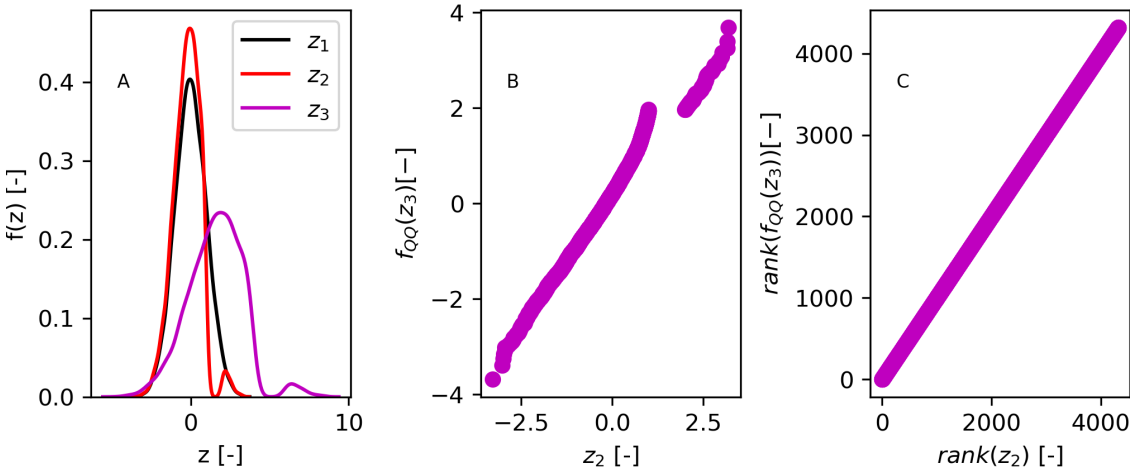

Figure 3.1: A) PDFs of z_1: as a reference dataset with $z_1 = \Phi(0,1)$; z_2: a subset of z_1, in which the values between $(1,2]$ are not sampled; z_3: an observation of z_2 with $z_3 = 2z_2 + 2$. B) The scatter plot between z_2 and the QQ-transformed z_3. C) The scatter plot between z_2 and the QQ-transformed z_3 in rank space.

Figure 3.2 shows simulations of PA using point values and/or order of point values to mimic the reference random field Figure 3.2A). The ensemble mean and variance of 100 realization are drifted by the included information: 1) 50 point values (white cross sign) as direct observations (Figure 3.2B and C) 2) order of 50 point values (white plus sign) as order information (Figure 3.2E and F) 3) 50 point values and the order of 50 different point values (Figure 3.2 H and I). The light and dark areas, e.g., high-K and low-K, in the plots of $\mu(F_Z)$ represent the simulated local features, which are generated based on the conditional points. These local features are shown as the dark areas, e.g., low variance and low uncertainty, in the plots of $\sigma^2(F_Z)$. The models with both point values and the order of point values (Figure 3.2H and I) have the largest light and dark area in the plot of $\mu(F_Z)$ and the largest dark area in the plots of $\sigma^2(F_Z)$ than the other two models with only a part of the information, which means a better result can be obtained using more information, although one is direct observation and another one is an observation of the order.

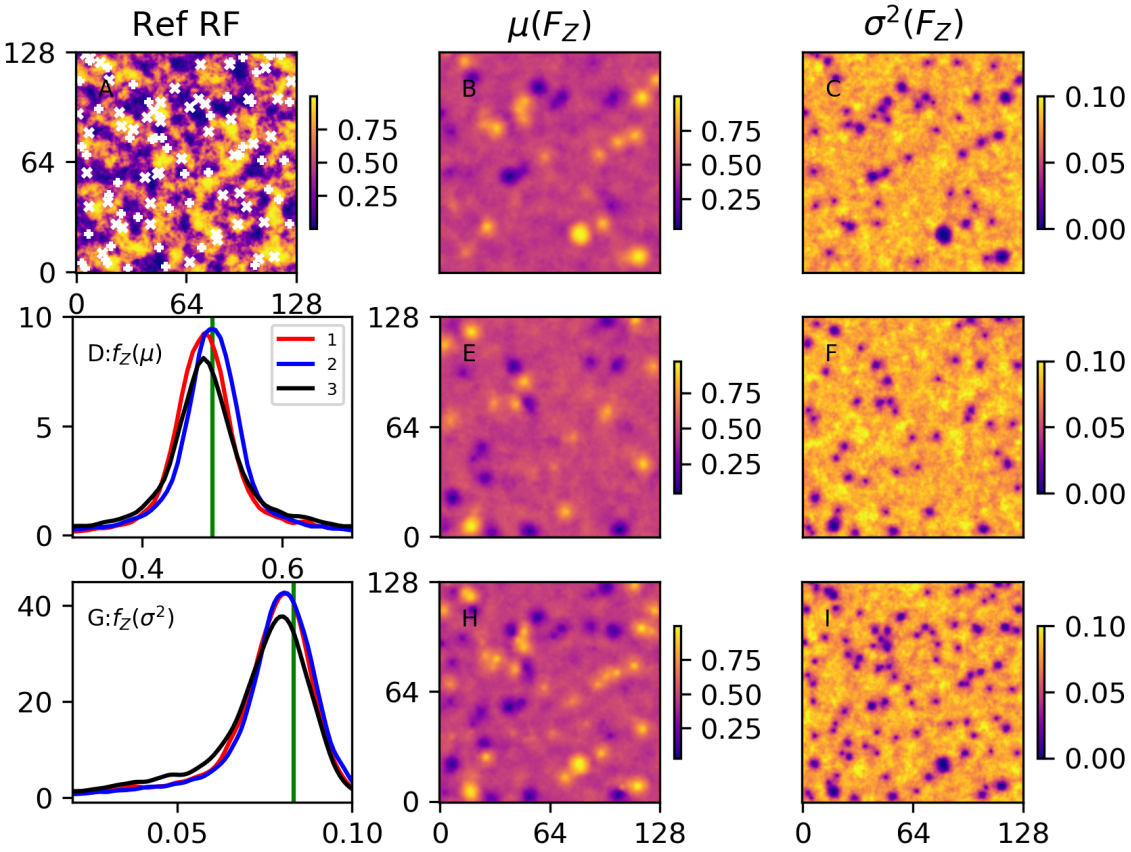

Figure 3.2: Conditional simulations of 100 realizations with point values (white cross sigh) and/or order of point values (white plus sigh). A) the reference random field. B) the ensemble mean and C) ensemble variance with 50 points values (case 1). E) the ensemble mean and F) ensemble variance with 50 order of point values (case 2). H) the ensemble mean and I) ensemble variance with 50 points values and 50 order of points values (case 3). D) The distribution function of the ensemble mean (green line: $\mu(F_Z)$) and G) variance (green line: $\sigma^2(F_Z)$) of each case.

Figure 3.2D and E present the density functions of the ensemble mean and ensemble variance. Both types of information have a part of the reference "truth". Therefore, the realizations containing both information can better represent the reference random field than the realizations with only one information, and the uncertainty of the simulated realizations is reduced. The realizations with both information have a flatter density function of mean values and a density function of variance with a large portion of the small variance.

3.3 Asymmetry Simulation Using V-transformation

A non-multi-Gaussian random field can be simulated using a non-monotonic transformation of a multi-Gaussian random field (Zinn and Harvey, 2003), e.g., the v-transformation in this thesis (Equation 2.32). The required parameters of the v-transformation and the underlying correlation function $\rho_{s,init}$ of the multi-Gaussian random field before the v-transformation can be estimated from the observations using the maximal likelihood method (Section 2.4.2). When m_c and k_c are known, an unconditional v-transformed non-multi-Gaussian random field can be simulated using the v-transformation f_v (Equation 2.32) as in Algorithm 2 (Li, 2010).

Algorithm 2: Unconditional simulation using v-transformation.

Result: RF_{out} with the required properties
initialization;
RF_g with the required $\rho_{s,init}$ is simulated ;
$RF_v \leftarrow f_v(RF_g)$;
$RF_{out} \leftarrow F_Z^{-1}(F_Z(RF_v))$ transform to the required F_Z;

The relationship between m_c, k_c and the spatial correlation before ($\rho_{s,init}$) and after the v-transformation (ρ_s) is non-injective, which means different combinations can generate random fields with the same spatial correlation (Gong *et al.*, 2013). Figure 3.3 shows the indicator plots of a different combination of the covariance and the v-transformation with an identical v-transformed spatial dependence structure. Different patterns of small values and large values can be found in the plot of multi-Gaussian field fields (g and $v : m_c = 55.0; k_c = 5.0$) and non-multi-Gaussian random fields.

Algorithm 2 is computationally efficient because the v-transformation is performed directly on a multi-Gaussian random field RF_g with the required $\rho_{s,init}$, which can be simulated using the spectral simulation method. One method to use the v-transformation for a conditional simulation is including the calculation of the conditional copula density in a sequential simulation (Li, 2010). This requires a multivariate integration on each point and is computationally costly.

Another method to extent Algorithm 2 to a conditional simulation is integrating an interpolation function into the phase-annealing algorithm. The updated random field with the required underlying covariance structure after the phase randomization (Equation 3.10) has a normal margin (Nur *et al.*, 2005). So, an additional transformation $\Phi \rightarrow f_v \rightarrow F_z \rightarrow \Phi'$ is needed to calculate the objective function Equation 3.12:

$$obj_1 = \frac{MSE(y_c(\mathbf{x_i}), f_v^*(y^*(\mathbf{x_i})))}{obj_{1,init}}, \tag{3.16}$$

and

$$f_v^*(y^*(\mathbf{x_i})) = \Phi^{-1}(F_z(f_v(y^*(\mathbf{x_i})))). \tag{3.17}$$

Then a conditional random field can be simulated using Algorithm 3.

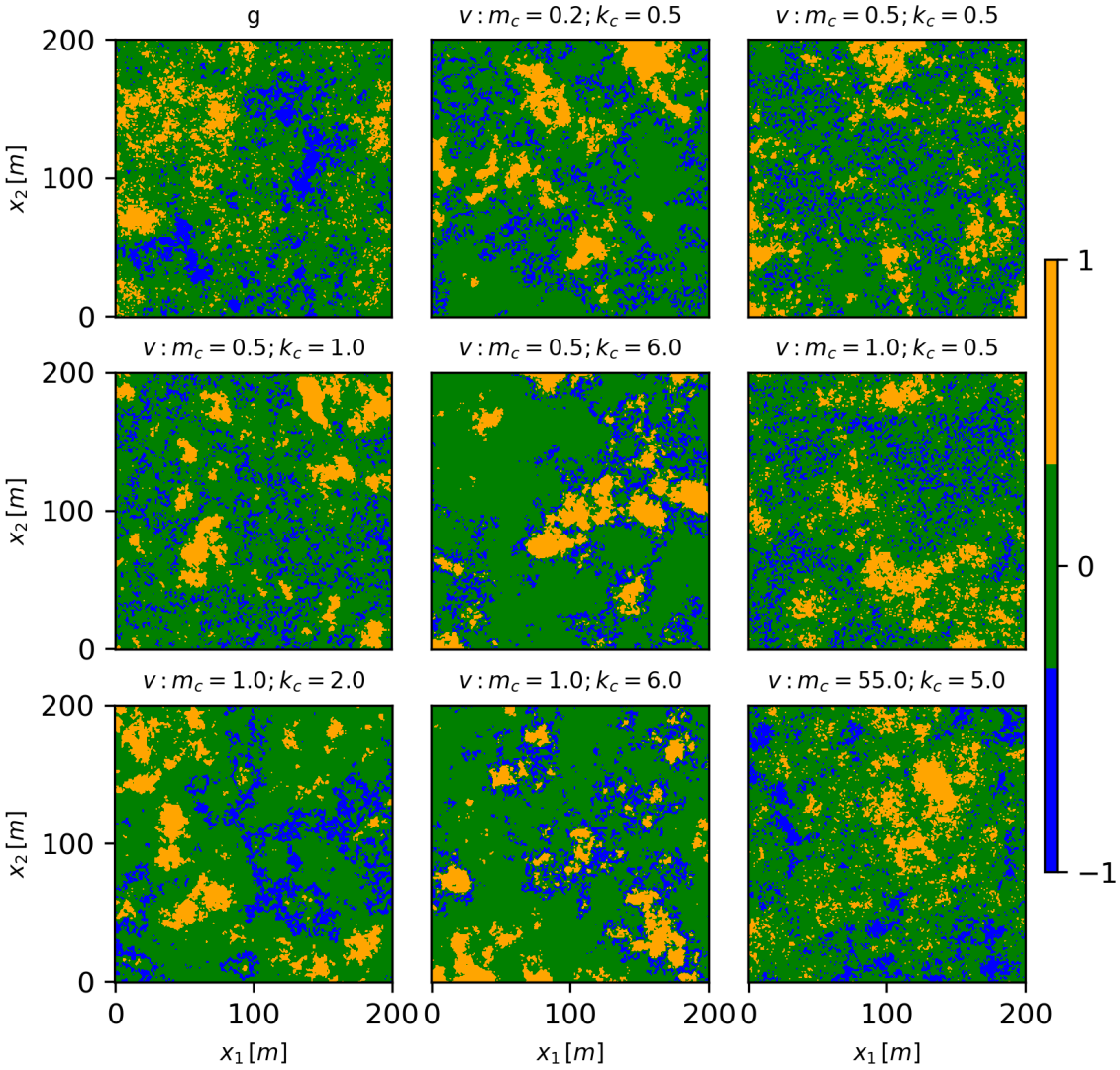

Figure 3.3: Indicator plots of unconditional simulation using the v-transformation with different m_c and k_c. All the random fields have an identical covariance structure.

Figure 3.4 shows the plots of conditional multi-Gaussian and v-transformed random fields using Algorithm 3. After including the 50 conditional points (white cross sigh), more similarities can be found between different plots than the unconditional results in Figure 3.3

Algorithm 3: Conditional simulation using V-transformation.

Result: RF_{out} with the required properties
initialization;
$f_v^*(y^*(\mathbf{x_i})) \leftarrow \Phi^{-1}(F_z(f_v(y^*(\mathbf{x_i}))))$;
RF with the required underlying $\rho_{s,init}$ as the initial status;
while $N_1 < N_{max,iter1}$ *and* $obj > obj_{tol}$ **do**
  $T \leftarrow T(T_0, N_1, N_{max,iter1})$ update annealing temperature;
  while $N_2 < N_{max,iter2}$ *and* $obj > obj_{tol}$ **do**
    $\psi^* \leftarrow random(0, 2\pi)$;
    $y^*(z(\mathbf{x})) \leftarrow$ update point values;
    $y_v^*(z(\mathbf{x})) \leftarrow f_v^*(y^*(z(\mathbf{x})))$
    $\mathbf{M}^*((y_v^*(x_i), y_v^*(x_j))) \leftarrow$ update the order matrix ;
    $obj^* \leftarrow$ update objective function;
    if $obj_{new} < obj_{best}$ **then**
      $obj, obj_{best} \leftarrow obj^*$;
    else if $random(0,1) \leq P_{accept}(obj, obj^*, T)$ **then**
      $obj \leftarrow obj^*$;
    else
      reject current perturbation;
    end
    $N_2 \leftarrow N_2 + 1$;
  end
  $N_1 \leftarrow N_1 + 1$;
end
$RF_{out} \leftarrow f_v(RF_{out})$;
$RF_{out} \leftarrow F_Z^{-1}(F_Z(RF_{out}))$ transform to the required F_Z ;

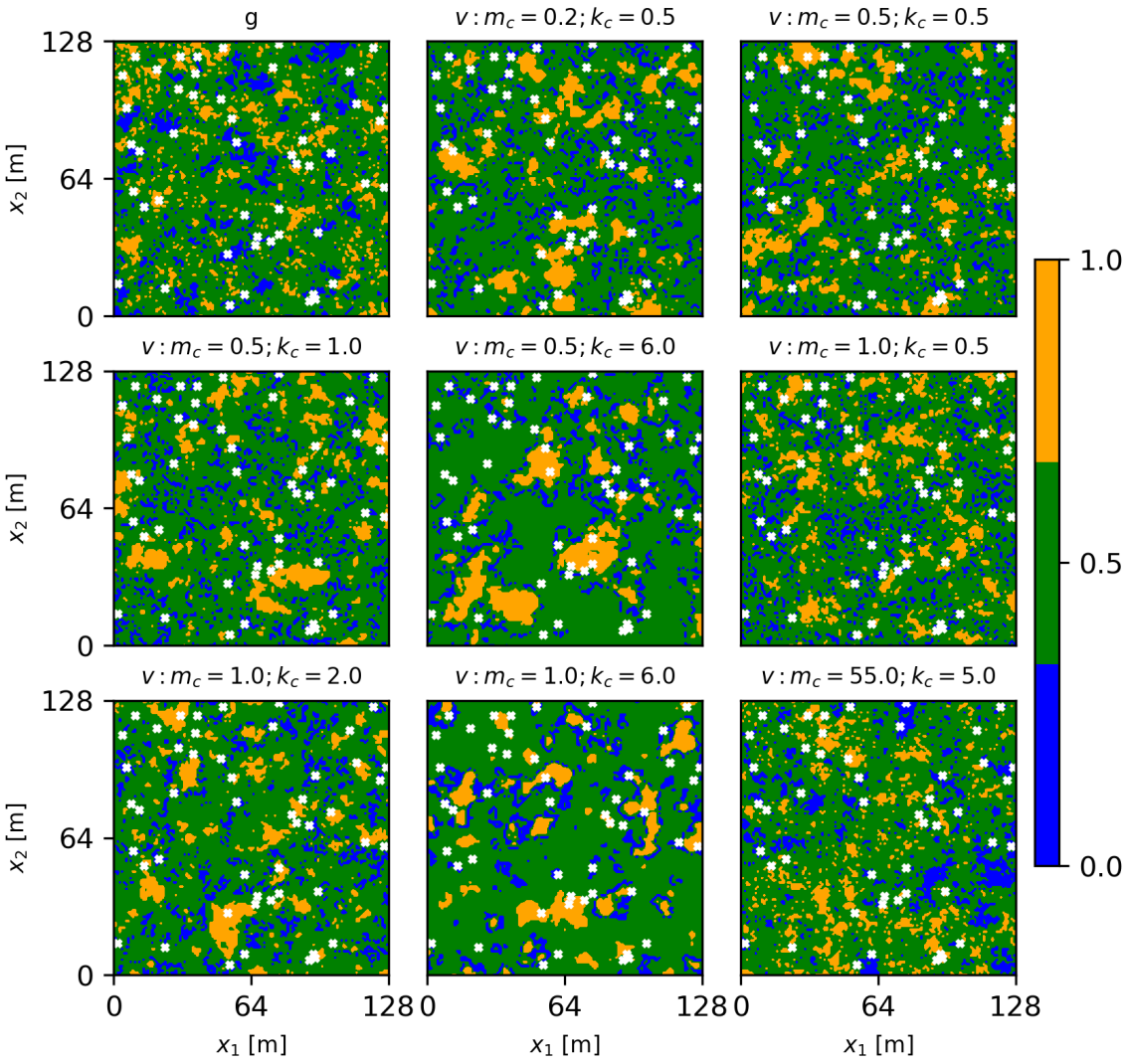

Figure 3.4: Indicator plots of conditional simulations using the v-transformation with different m_c and k_c. 50 conditional points (white cross sigh) are included in the simulation. All the random fields have an identical covariance structure.

3.4 FFT-Asymmetry

The requirement of the degree of Gaussianity can also be included in PA using the empirical asymmetry (Equation 2.27). Because the number of data pairs in Equation 2.27 is proportional to the square of the number of simulated points n, the updating of the asymmetry in PA iterations is computationally expensive, especially for a three-dimensional simulation. Possible methods to reduce the computational cost are sampling a small subset

of the dataset to calculate the asymmetry or storing the indices of the data pairs in memory and load them while updating the asymmetry. However, neither of these two approaches is efficient for a complex three-dimensional simulation, which requires a large number of iterations. Therefore, a Fast Fourier Transform (FFT) based algorithm on a regular grid (Marcotte, 1996) is used to reduce the complexity required to update the asymmetry from n^2 to $n\ln(n)$. The asymmetry (Equation 2.27 and Equation 2.28) of a set of values $z(\mathbf{x})$ in the copula space can be calculated by:

$$A = \frac{\mathbf{TF^{-1}}\left\{\begin{array}{c}\overline{F_2}\cdot F_1 - 0.5\cdot\overline{F_2}\cdot F_I - 2\cdot\overline{F_1}\cdot F_1 + \overline{F_1}\cdot F_2 \\ +0.75\cdot\overline{F_1}\cdot F_I - 0.5\cdot\overline{F_I}\cdot F_2 + 0.75\cdot\overline{F_I}\cdot F_1 - 0.25\cdot\overline{F_I}\cdot F_I\end{array}\right\}}{\mathbf{TF^{-1}}\left\{\overline{F_I}\cdot F_I\right\}}, \tag{3.18}$$

in which $F_1 = \mathbf{TF}(F_Z(\mathbf{x}))$, $F_2 = \mathbf{TF}(F_Z(\mathbf{x})^2)$, F_I is the Fourier transform of a logical matrix, in which 1 and 0 indicate a location with or without an observation and $\overline{(\cdot)}$ is the complex conjugate.

Equation 3.18 can be simplified as

$$A = \frac{\mathbf{TF^{-1}}(\overline{\tilde{F}_2}\cdot\tilde{F}_1 + \overline{\tilde{F}_1}\cdot\tilde{F}_2)}{\mathbf{TF^{-1}}(\overline{F_I}\cdot F_I)}, \tag{3.19}$$

in which $\tilde{F}_1 = \mathbf{TF}(F_Z(\mathbf{x}-0.5))$, $\tilde{F}_2 = \mathbf{TF}((F_Z(\mathbf{x})-0.5)^2)$. So, the asymmetry in Equation 2.27 and Equation 2.28 describes the cross-correlation between $\mathbf{FT}(F_Z(\mathbf{x}))$ and $\mathbf{FT}(F_Z(\mathbf{x})^2)$ and it is zero for a multi-Gaussian random field.

Equation 3.19 can be rewritten as

$$A = \frac{\mathbf{TF}^{-1}\left(2\cdot<\tilde{F}_1, R_{xx}(\tilde{F}_1)>\right)}{\mathbf{TF^{-1}}(\overline{F_I}\cdot F_I)}, \tag{3.20}$$

in which $<\cdot,\cdot>$ is the inner product operator and R_{xx} is the auto-correlation function. Equation 3.20 connects the spatial structure $A(F_Z(\mathbf{x}))$ in the spatial coordinates and the spatial structure $R_{xx}(\mathbf{FT}(F_Z(\mathbf{x})-0.5))$ in the frequency space. In the case of a multi-Gaussian random field, there is $<\tilde{F}_1, R_{xx}(\tilde{F}_1)>=0$. This means $\tilde{F}_1$ is orthogonal to $R_{xx}(\tilde{F}_1)$. A detailed derivation and simplification of Equation 3.18-3.20 can be found in Appendix A.

A conditional non-multi-Gaussian random field can be simulated using Algorithm 4 with Equation 3.19. Figure 3.5 shows the ensemble measures of 50 realizations of one simulation with 50 point values and 50 order-of-point values used in conditioning.

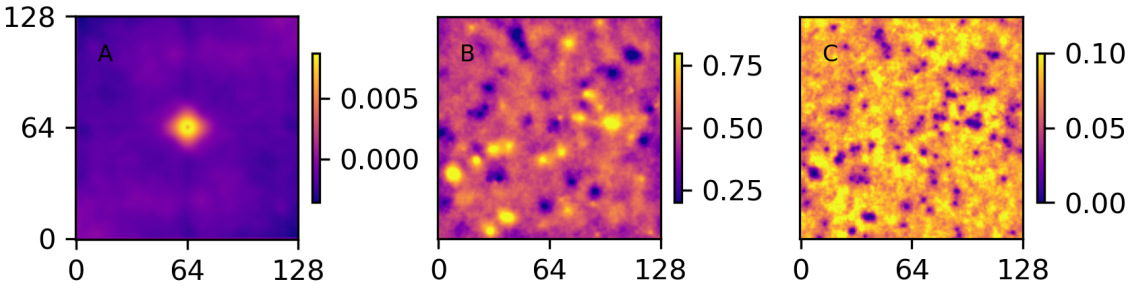

Figure 3.5: Ensemble properties of 50 realizations of PA simulations with 50 point values, 50 order of point values, and FFT-asymmetry as constraints. A) The ensemble mean value of the asymmetry map. B) The ensemble mean value and C) ensemble variance.

Algorithm 4: Phase-annealing with FFT-asymmetry.

Result: RF_{out} with the required properties
initialization;
RF with the required variogram as the initial status;
while $N_1 < N_{max,iter1}$ *and* $obj > obj_{tol}$ **do**
  $T \leftarrow T(T_0, N_1, N_{max,iter1})$ update annealing temperature;
  while $N_2 < N_{max,iter2}$ *and* $obj > obj_{tol}$ **do**
    $\psi^* \leftarrow random(0, 2\pi)$;
    $y^*(z(\mathbf{x})) \leftarrow$ update point values;
    $\mathbf{M}^*((y^*(x_i), y^*(x_j))) \leftarrow$ update the order matrix ;
    $A(\mathbf{h})_{new}(F_Z(\mathbf{x}), F_Z(\mathbf{x}+\mathbf{h})) \leftarrow$ update asymmetry ;
    $obj^* \leftarrow$ update objective function;
    if $obj_{new} < obj_{best}$ **then**
      $obj, obj_{best} \leftarrow obj^*$;
    else if $random(0,1) \leq P_{accept}(obj, obj^*, T)$ **then**
      $obj \leftarrow obj^*$;
    else
      reject current perturbation;
    end
    $N_2 \leftarrow N_2 + 1$;
  end
  $N_1 \leftarrow N_1 + 1$;
end
$RF_{out} \leftarrow$ backward transformation ;
$RF_{out} \leftarrow F_Z^{-1}(F_Z(RF_{out}))$ transform to the requried F_Z ;

3.5 Computational Aspects of Phase-Annealing

Calculation of $F_1(\mathbf{x})$ and $F_2(\mathbf{x})$ There are two different options to reduce the computational cost of the calculation of $F_1(\mathbf{x})$ and $F_2(\mathbf{x})$. Because $F_Z(\mathbf{x})$ is defined in the real space $\mathbb{R}^n$, a real number FFT can be used to replace the normal FFT. Due to the symmetry of the Fourier coefficients of real data, only half of the Fourier coefficients are needed to be calculated. This means only half of the points are needed to be calculated in Equation 3.19.

Another option is that two real numbers ($F_1(\mathbf{x}) = \mathbf{FT}(f_1)$ and $F_2(\mathbf{x} = \mathbf{FT}(f_2))$ FFT can be calculated using one complex number FFT on $f_{12}(\mathbf{x}) = f_1(\mathbf{x}) + i \cdot f_2(\mathbf{x})$. Then the k-th term of the FT for a random field with n points are

$$F_{1,k}(\mathbf{x}) = \frac{1}{2}[\mathbf{RE}(F_{12,k}(\mathbf{x})) + \mathbf{RE}(F_{12,n-k}(\mathbf{x}))] + i \cdot \frac{1}{2}[\mathbf{Im}(F_{12,k}(\mathbf{x})) - \mathbf{Im}(F_{12,n-k}(\mathbf{x}))], \tag{3.21}$$

$$F_{2,k}(\mathbf{x}) = \frac{1}{2}[\mathbf{Im}(F_{12,k}(\mathbf{x})) + \mathbf{Im}(F_{12,n-k}(\mathbf{x}))] - i \cdot \frac{1}{2}[\mathbf{Re}(F_{12,k}(\mathbf{x})) - \mathbf{Re}(F_{12,n-k}(\mathbf{x}))]. \tag{3.22}$$

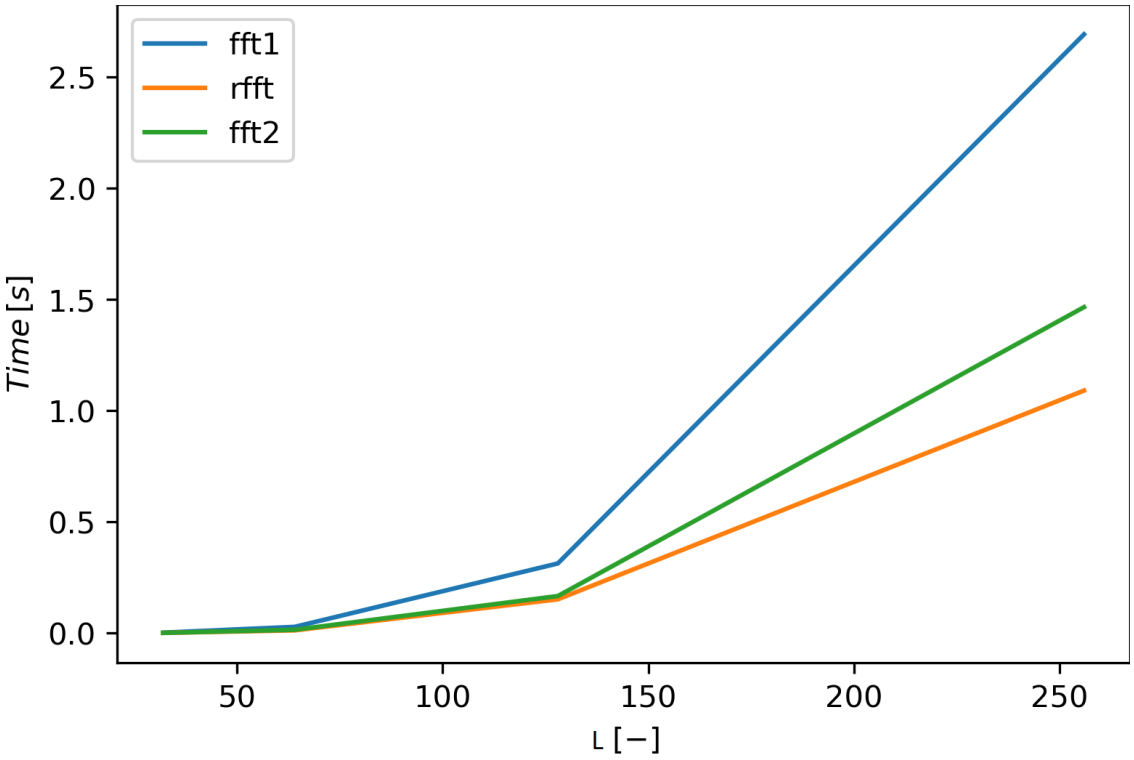

Figure 3.6: Computational cost of normal FFT (fft1), real value FFT (rfft) and combine two FFT in one (fft2) on three-dimensional fields with a domain size $\mathbf{L} = (L;L;L)$.

The computational costs of these two methods are briefly tested and compared with the normal FFT in Figure 3.6. Both methods can improve the performance significantly.

Update of the Distribution Function For a random field with a large number of points, $F_Z(\mathbf{x})$ can be approximated by $F_\Phi^{-1}(y(\mathbf{x}), \mu, \sigma^2)$, where F_Φ^{-1} is the percent point function (ppf) or inverse function of a normal distribution with mean μ and variance σ^2.

Selection of the Range of the Random Phases The intensity of the perturbation in PA iterations depends on the selection of the phase to be randomized, which can be classified as high frequency and low frequency. Figure 3.7 shows how a random field changes after shifting a low frequency and a high frequency with the same degree. Low frequencies contain the information of the global structure of the RF and lead to a large variation after shifting. High frequencies contain local detail information and lead to a small variation after shifting.

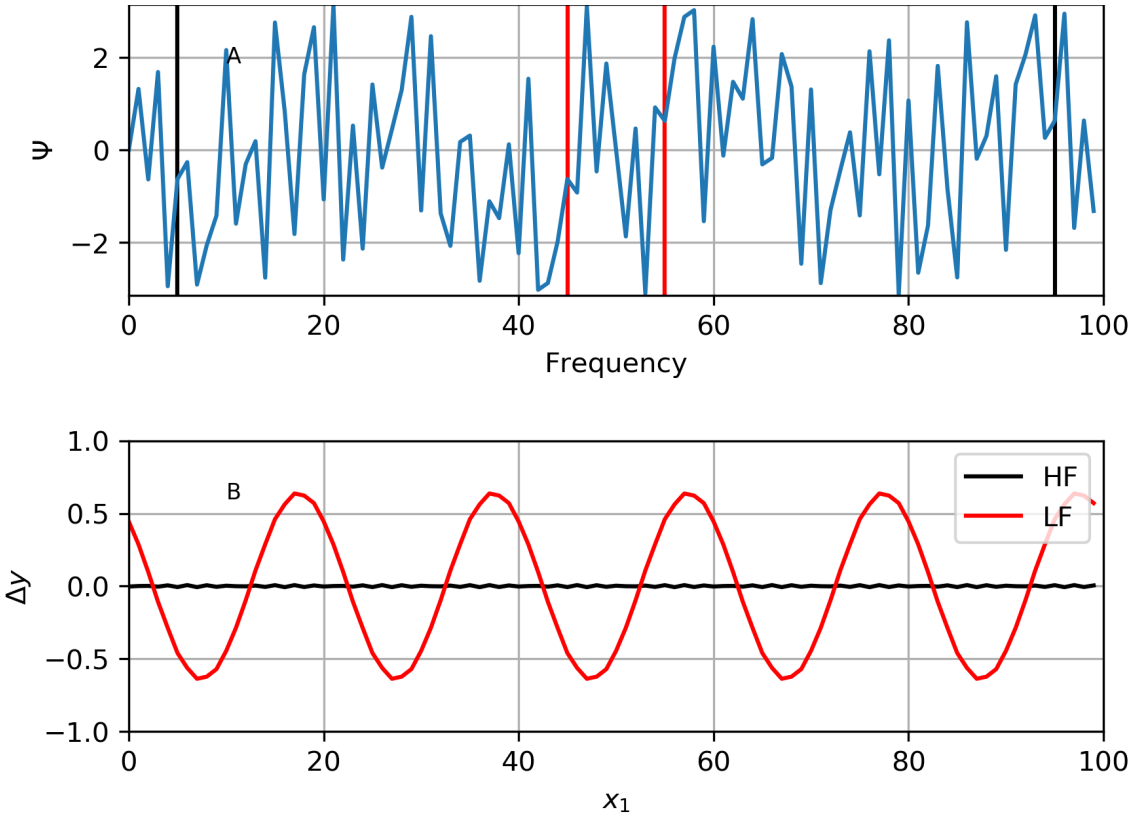

Figure 3.7: A low-frequency and a high-frequency A) on the frequency domain and B) their influences on the RF after a random phase shift.

Figure 3.8 shows the averaged rejection ratio and the objective function of 10 simulations versus the annealing temperature T. The best selection of the phase range to be randomized **rn** depends on the required domain size **L** to be simulated. A large **rn** has a small rejection ratio but also an objective function with a small descend gradient. In this case, a dynamic phase range can be used (Figure 3.8 $\mathbf{L} = 1000^2$). A small phase range ($\mathbf{rn} = 40^2$) is used when the annealing temperature is high to get a large gradient of the objective function. Then a large phase range ($\mathbf{rn} = 120^2$) is used when the annealing temperature is low to get a smaller objective function.

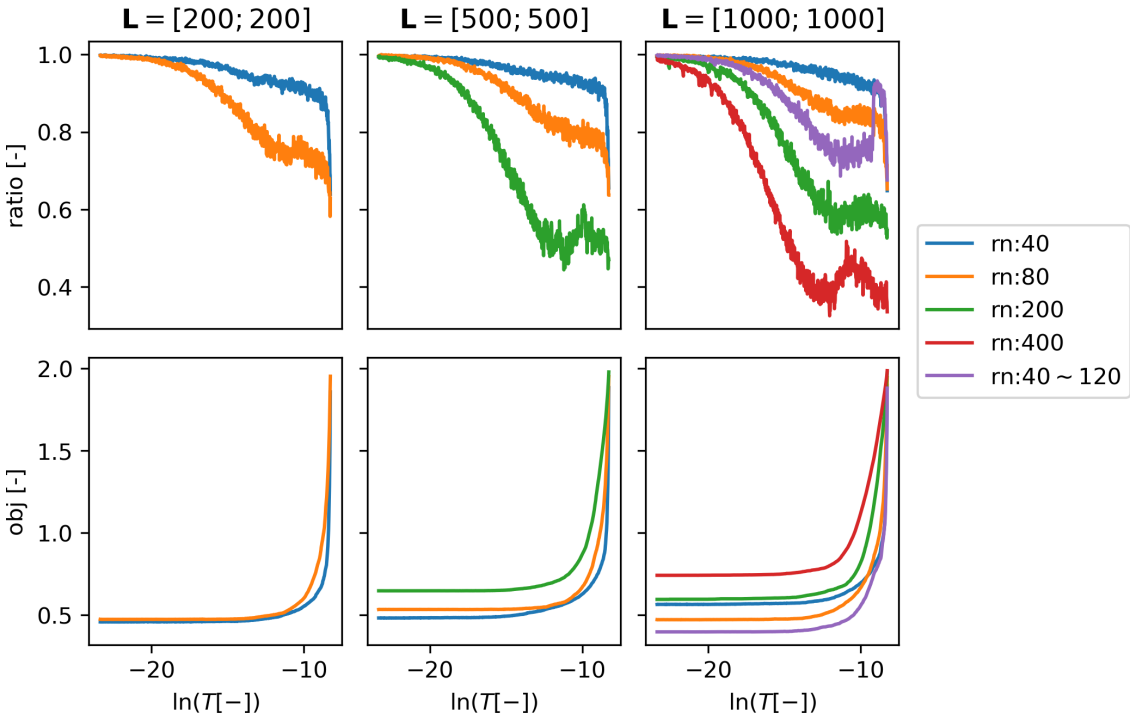

Figure 3.8: Average ratio of rejection of a perturbation (first row) and objective function (second row) of 10 simulations versus annealing temperature with different domain size $\mathbf{L} = 200^2; 500^2; 1000^2$ of RF with a correlation structure $1.0Exp(10)$ and different fixed phase range $\mathbf{rn} = 40^2; 80^2; 200^2; 400^2$ and dynamic phase range $\mathbf{rn} = (40 \sim 120)^2$ for the phase randomization.

Selection of the Input Domain Size Figure 3.9 shows different levels of the domain size to calculate in PA. To simulate a RF with a domain size $\mathbf{L} = (L_1; L_2; ...; L_n)$ (the red box in Figure 3.9A), a large input RF with the required correlation structure and the domains size $\mathbf{L_{in}} = 2 \cdot \mathbf{L} - 1$ (blue box in Figure 3.9A) is simulated at first to reduce the boundary effect. This input domain is transformed to the Fourier space as in Figure 3.9B and the range of the phases to be randomized is selected (black box in Figure 3.9B), which means only the phase and amplitude in this range are needed during PA. For a large-scale

three-dimensional simulation, it can be taken out during PA to reduce the memory usage and the CPU time for indexing. In the case with asymmetry, only the point in the red box but not the whole input RF is updated. An additional zero-padding is performed to calculate Equation 3.18. RF is normally extended to $\mathbf{L_{ext}} = 2 \cdot \mathbf{L}$, which is the smallest even number large than $2 \cdot \mathbf{L} - 1$.

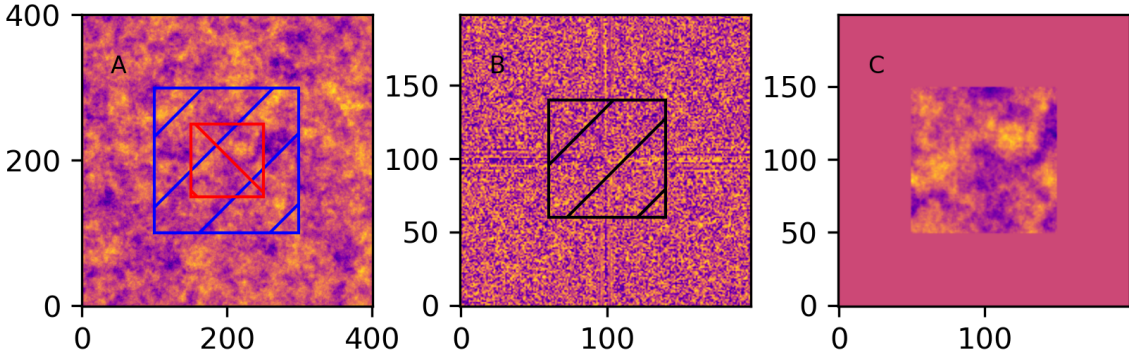

Figure 3.9: A) Input RF in the blue box as the initial status of the annealing iterations and required domain size in the red box; B) Range for the phase randomization in the Fourier space (black box); C) Domain size with a zero padding to update the asymmetry in the annealing iterations.

A specific issue in the three-dimensional K simulation is that the borehole observations normally have a high resolution in the vertical direction. For a conditional simulation using these observations, a random field with a shorter length and finer resolution in the vertical direction than in the horizontal direction is simulated. Figure 3.10 shows the influence of the spatial distribution of the conditional points on the objective function. The test case 2 has conditional points with a dense spatial distribution. After several PA iterations, its objective function has a smaller chance to reach a small value. In this case, the input random fields with the domain size $\mathbf{L_{in}} = 2 \cdot \mathbf{L} - 1$ cannot successfully be conditioned on the observations due to the lack of the variability in Fourier space. To solve this issue, a larger input random field or a zero-padding on the input random field is needed.

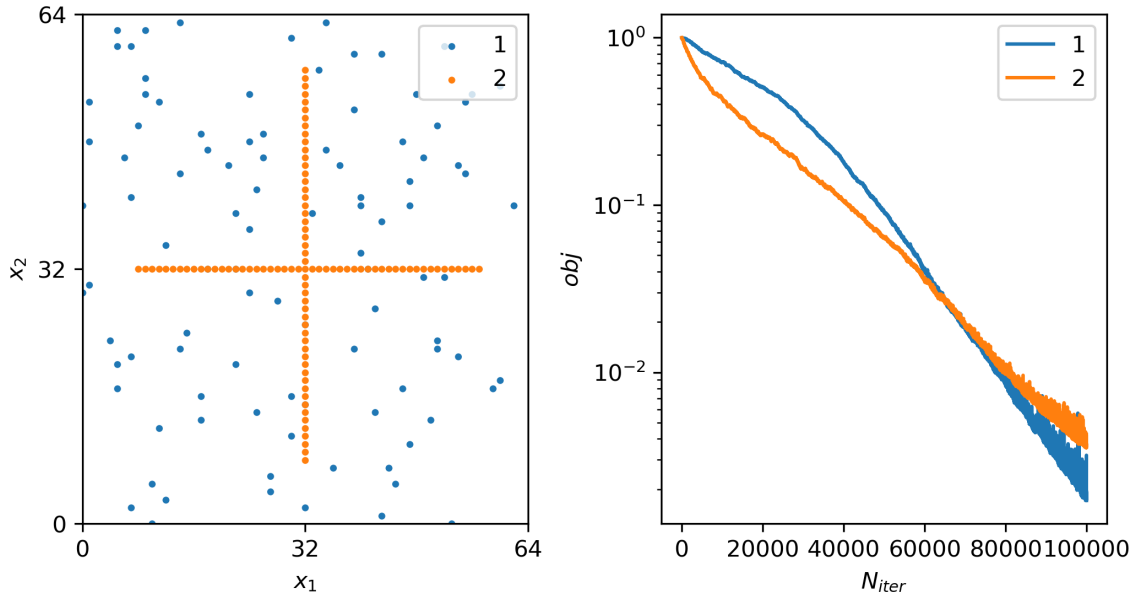

Figure 3.10: A) Conditional points with a dispersed spatial distribution (1) and a dense spatial distribution (2). B) The corresponding plot of the objective function versus the number of iteration of 50 simulations in PA.

Dimensionality reduction of the inverse FFT In Equation 3.19, one inverse FFT is needed to obtain the current value of asymmetry. Normally, only the asymmetries in the axis directions are used in the objective function, which means only a part of the asymmetry map is needed. According to the projection slice theorem (Deans, 1983), the output of the Radon Transform of a two-dimensional random field corresponds to a line in the Fourier space. This is equalized to the inverse Fourier Transform of the time-reversed relationship between the forward Fourier Transform and the inverse Fourier Transform. Therefore, a two-dimensional inverse Fourier Transform is transferred to the Radon Transform plus some one-dimensional inverse Fourier Transform. This method can be used as an extension in the future.

Additional benefits from computational libraries All geostatistical algorithms in this thesis are implemented in Python. Due to the high computational cost after including asymmetry in the objective function, a comparison between the performance of different computational libraries for the FFT calculation and matrix operations is shown in Appendix B. This analysis was necessary to detect the best combination based on the available computational resources.

Patterns in the Fourier Space Phase Annealing is a combinatorial optimization to simulate the distribution of the phase angles in the Fourier space according to the conditional constraints. Therefore, an investigation into the structure of the phase angle is a possible way to reduce the computational cost of the simulations with asymmetry in the future.

Figure 3.11 shows the ensemble mean μ and variance σ^2 of the phase angle ψ of simulations without (A and B) and with (C and D) asymmetry. A stronger pattern can be found in the simulation without asymmetry than in the simulation with asymmetry. This proves that more variability is enforced in the simulated K-field after including asymmetry in the simulation by driving the output K-field away from being multi-Gaussian. Even so, patterns can be found in both simulations. Therefore, a further study on these patterns of the phase angle is interesting in the future.

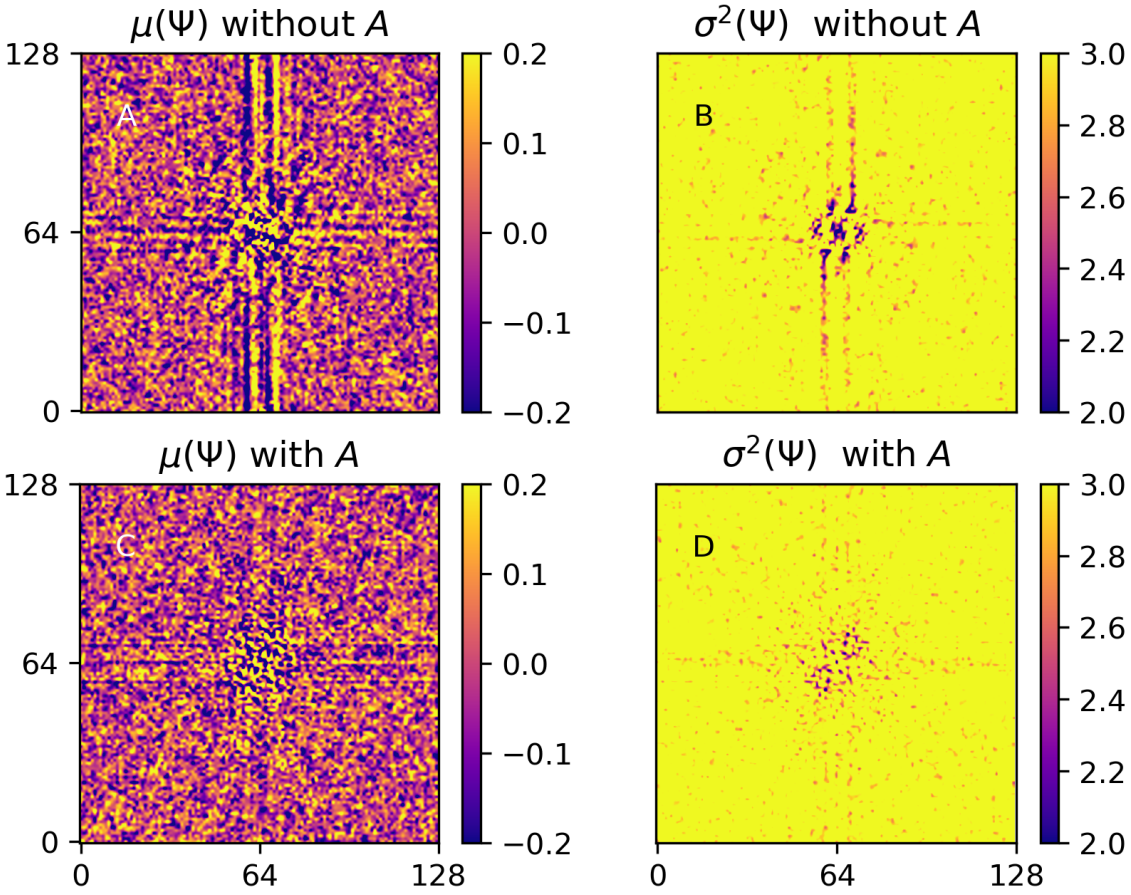

Figure 3.11: Ensemble mean μ and variance σ^2 of the phase angle ψ of simulations without (A and B) and with asymmetry (C and D).

3.6 Summary of Chapter 3

In this chapter, phase-annealing-based methods have been introduced (Section 3.2) as a flexible tool for conditional non-multi-Gaussian geostatistical simulations. Various types of information, including, but not limited to, the point value, the order-of-the-point value, the two-point spatial dependence variogram, and the third-order spatial dependence asymmetry, were integrated into the simulation to include as much information as possible from the observations (Section 3.2.2).

There are two alternative ways to include the copula-based non-Gaussianity in PA. When the v-copula introduced in Chapter 2 is assumed as a theoretical copula model, the non-Gaussianity can be described using a non-monotonic v-transformation (Algorithm 3). The non-Gaussianity can be also described directly using the copula asymmetry (Algorithm 4). To reduce the computational complexity, a Fourier Transform based description of asymmetry (section 3.4) and further methods (Section 3.5) have been introduced.

Normally the estimation of the empirical asymmetry extracts additional information from the existing observations. Therefore, it does not require more observations than the standard second-order stationary geostatistics. A possible extension in the future could be to use more readily available geophysical measurements for the estimation of asymmetry on the field scale.

One disadvantage of the method is the high computational cost to recompute the asymmetry in the phase-annealing method. In standard multi-Gaussian conditional simulations, efficient techniques such as the sequential Gaussian simulation or the method of smallest modification (Deutsch and Cockerham, 1994) can be used. Unfortunately, these methods cannot be conditioned on asymmetry information in PA. Conversely, the phase-annealing method does not simply allow fixing point values as the method of smallest modification. The phase spectrum is randomized and an iterative search algorithm seeks for the best phase minimizing the overall objective function, including the point values. The phase-annealing approach is accelerated by FFT techniques and runs the Monte Carlo simulations in parallel on a high-performance cluster (HPC), but the reduction of the computational costs for three-dimensional simulations is still interesting future work. Further improvements might be achievable by using a wavelet-based local update method to replace the FFT-based global update phase randomization method in phase-annealing.

Chapter 4

Travel-Time Based Evaluation of Macrodispersion

To evaluate the effects of including non-multi-Gaussian spatial dependence in the K simulation on the solute transport, a cell-centered Finite Volume method is used to simulate groundwater flow and particle-tracking random-walk simulations of solute transport are performed using the simulated K fields. In Section 4.1 the theoretical background of the particle-tracking random-walk simulation method is introduced. Preliminary simulation results of the test scenarios are analyzed in Section 4.2.

4.1 Particle-Tracking Random-Walk Simulation

The velocity field meets Darcy's law and the continuity equation without sources and sinks:

$$n\mathbf{v} = -K\nabla h_w \tag{4.1}$$

$$\nabla \cdot (n\mathbf{v}) = 0 \tag{4.2}$$

in which $\mathbf{v}$ is the linear velocity, n denotes porosity, and $h_w(\mathbf{x})$ is the hydraulic-head field. In the application of this thesis, fixed hydraulic heads at the in- and outlet faces of a rectangular domain are assumed, and no-flow conditions at all other boundaries. Then the semi-analytical method of Pollock (1988) with graphics processing unit (GPU) acceleration for particle tracking to address advective transport is used, amended by a random walk to account for local dispersion (Tompson and Gelhar, 1990) using the standard Scheidegger parameterization (Scheidegger, 1961):

$$\mathbf{D} = (D_p + \alpha_t |\mathbf{v}|)\,\mathbf{I} + \frac{\mathbf{v} \otimes \mathbf{v}}{|\mathbf{v}|}\,(\alpha_\ell - \alpha_t) \tag{4.3}$$

in which D_p is the pore-diffusion coefficient, α_ℓ and α_t are the local longitudinal and transverse dispersivities, respectively, $\mathbf{I}$ is the identity matrix, and $\otimes$ is the matrix product. In the simulations of this thesis, the local dispersion in the x_ℓ-direction of mean flow is neglected and the breakthrough of particles at observation planes perpendicular to x_ℓ is analyzed.

A standard approach of characterizing solute transport in heterogeneous domains is by analyzing the temporal development of the spatial moments of solute plumes, in which the rate of change of the vector of first moments yields the plume-effective velocity, and half the rate of change of the matrix of second-central moments yields an operational definition of the macrodispersion tensor (e.g. Freyberg, 1986; Gelhar and Axness, 1983, among others). An alternative approach is to consider the travel-time (τ) distribution of solute particles from one observation plane to the next in the direction of the mean flow. The Lagrangian velocity U_ℓ of a single particle in the main flow direction between two observation planes can be calculated from the pathway travel-time τ by:

$$U_\ell = \left(\frac{\partial \tau}{\partial x_\ell} \right)^{-1}. \tag{4.4}$$

Within the travel-time framework, the macroscopic longitudinal dispersion coefficient D_ℓ can be evaluated by matching the statistical properties of the travel-time distribution (Dagan *et al.*, 1992).

In this thesis, two different types of the dispersion coefficient differing in the order of taking moments and ensemble averages (Kitanidis, 1988; Dentz *et al.*, 2000) are distinguished: The *effective* (eff) dispersion coefficient samples the travel-time variability of many particles introduced at the same location in a single realization and subsequently averages over all realizations within the ensemble:

$$D_{\ell,eff} = \langle D_{\ell,i} \rangle = \left\langle \frac{1}{2} \cdot \frac{\partial \sigma^2_{\tau_i}}{\partial x_\ell} \cdot U^3_{\ell,i} \right\rangle, \tag{4.5}$$

whereas the *ensemble* (ens) dispersion coefficient is derived from first merging the travel times of all N realizations ($\tau_{ens} = \bigcup_{i=1}^{N} \tau_i$) and then evaluating the statistical moments:

$$D_{\ell,ens} = \frac{1}{2} \cdot \frac{\partial \sigma^2_{\tau_{ens}}}{\partial x_\ell} \cdot U^3_{\ell,ens}. \tag{4.6}$$

That is, the effective longitudinal dispersion measures the average longitudinal spread of solute plumes originating from a point injection as observed in single realizations, which is important for solute mixing (Cirpka and Kitanidis, 2000), whereas the ensemble dispersion includes the uncertainty in the mean arrival time among the different realizations.

In analogy to the longitudinal dispersion coefficient, an effective and an ensemble Lagrangian velocity are defined, depending on the order of taking the inverse of the derivative of travel time with x_ℓ and the ensemble average:

$$U_{\ell,eff} = \langle U_{\ell,i} \rangle = \left\langle \left(\frac{\partial \tau_i}{\partial x_\ell} \right)^{-1} \right\rangle, \tag{4.7}$$

$$U_{\ell,ens} = \left(\frac{\partial \langle \tau_{ens} \rangle}{\partial x_\ell} \right)^{-1} \tag{4.8}$$

The variability of the travel-time increments $\Delta\tau_i$ causes the effective Lagrangian velocity $U_{\ell,eff}$ to be considerably larger than the ensemble Lagrangian velocity $U_{\ell,ens}$.

4.2 Evaluation of the Solute Transport Characteristics

To show the influence of the degree of non-Gaussianity of K on the particle-tracking random-walk simulations, test scenarios with K fields of various geostatistical models (Table 4.1) are simulated and their effects on solute transport are modeled (Table 4.2).

2×10^3 particles are injected at $\mathbf{x_{inj}} = (40\,[m]; 10\,[m]; 20\,[m])$ as a conservative tracer at the beginning of the simulation. Two control parameters are used here for comparison, i.e., the marginal distribution and the degree of non-Gaussianity expressed in the measure of asymmetry. The marginal distribution M_1 (homogeneous models g_1, v_{11} and v_{12}) has a geometric mean $\mu = 9.54 \times 10^{-5}$ [m/s] and a small variance $\sigma^2(\ln(K[m/s]) = 0.39$. The marginal distribution M_2 (heterogeneous models g_2, v_{21} and v_{22}) has a geometric mean $\mu = 4.29 \times 10^{-5}$ [m/s] and a large variance $\sigma^2(\ln(K[m/s]) = 4.41$. Besides the marginal distribution, three different types of asymmetry are used. Models g_0 and g_1 are two multi-Gaussian models and models v_{11}-v_{22} are non-multi-Gaussian models with positive asymmetry (v_{11} and v_{12}) or negative asymmetry (v_{21} and v_{22}). The non-multi-Gaussian models are simulated using a v-transformation with parameters $m_c = 0.2$ and $k_c = 3.0$ (Section 3.3). These parameters are selected to show the influence of the degree of Gaussianity on the solute transport characteristics. In practice, parameters of the v-transformation can be estimated from the data using a theoretical v-copula model (Section 2.4.2). For each model, 200 realizations are generated.

Table 4.1: Geostatistical models of the test scenarios of the partial-tracking random-walk method.

	g_1	g_2	v_{11}	v_{12}	v_{21}	v_{22}
Marginal Distribution	M1	M2	M1	M1	M2	M2
Multi-Gaussian	✓	✓				
Asymmetry [-]	0	0	+	-	+	-

Table 4.2: Configurations of the test scenarios of K simulation and particle-tracking simulation.

Simulation of K-Field

Vertical correlation length	l_v	$[m]$	1.0
Horizontal correlation length	l_h	$[m]$	10.0
Domain size	$\mathbf{L}$	$[m]$	$(80; 150; 40)$
Grid Spacing	$\Delta\mathbf{x}$	$[m]$	$(1.0; 1.0; 0.10)$

Particle-Tracking Random-Walk Simulation

Mean hydraulic gradient	i	[-]	8×10^{-3}
Porosity	n	[-]	0.35
Transverse dispersion coefficient	D_t	$[m^2/s]$	1×10^{-8}
Number of particle	N_p	[-]	2×10^3
Injection location	$\mathbf{x_{inj}}$	$[m]$	$(40; 10; 20)$

Figure 4.1 shows indicator plots (Section 2.3.2) of the cross sections of one realization from model g_1 and g_2 (Figure 4.1A), v_{11} and v_{21} (Figure 4.1B), and v_{21} and v_{22} (Figure 4.1C). Different degrees of Gaussianity are presented as a different arrangement of large ($F_Z(\mathbf{x}) > 0.9$), middle ($0.1 < F_Z(\mathbf{x}) \leq 0.9$), and small ($0 < F_Z(\mathbf{x}) \leq 0.1$) quantiles of K values in the plots. The multi-Gaussian field (Figure 4.1A) has a similar shape for large values and small values. In contrast, the non-multi-Gaussian random field with a positive asymmetry (Figure 4.1B) has large blobs of large values and small blobs of small values and the one with a negative asymmetry A (Figure 4.1C) has large blobs of small values and small blobs of large values.

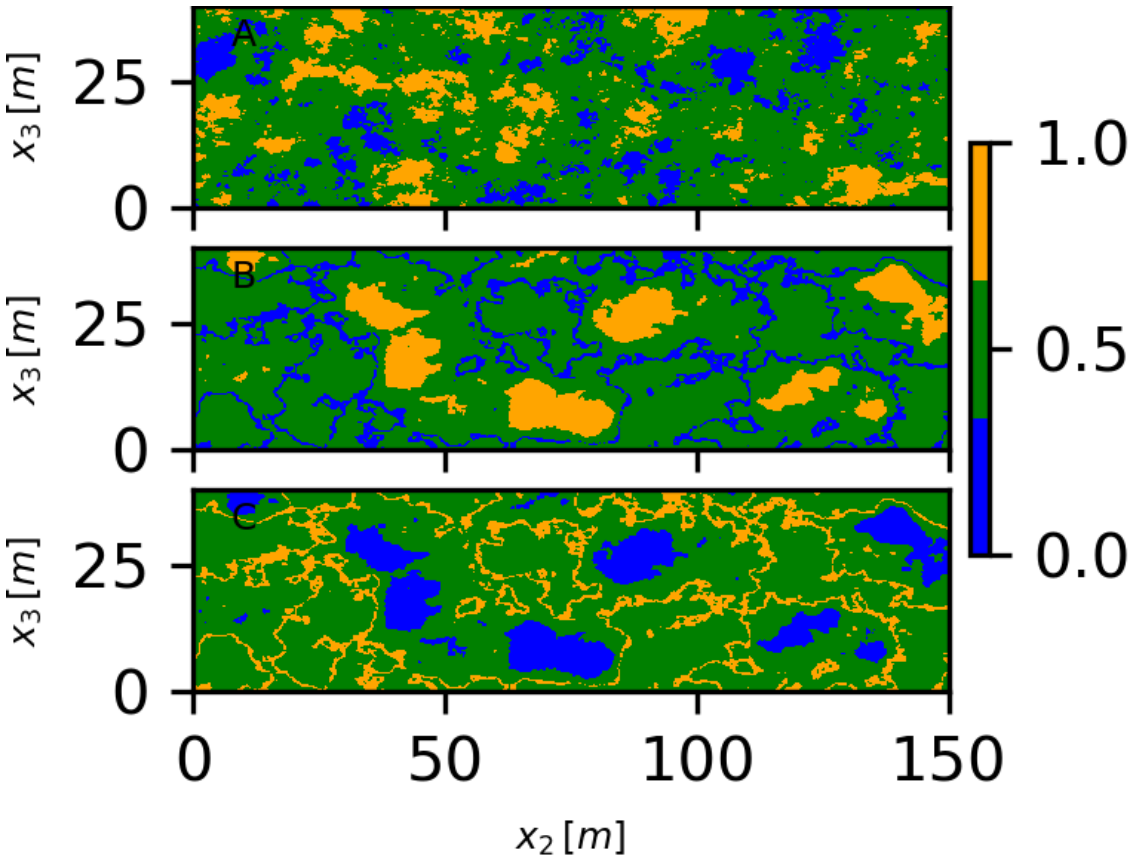

Figure 4.1: Indicator plots of cross sections of one realization of the simulated K fields at $x_1 = 40\,[m]$. A) One multi-Gaussian and B) non-multi-Gaussian ($m_c = 0.2$, $k_c = 3.0$) K fields with same parameters but positive asymmetry and C) negative asymmetry.

Numerical tracer tests are simulated on K-field by particle-tracking random-walk and analyzed by the particle travel time τ. The density functions of the log-travel time $\ln(\tau[s])$ of different models are presented in Figure 4.2. The homogeneous models (marginal distribution type M_1) have distributions with a small range than the heterogeneous models (marginal distribution type M_2). The influence of the included asymmetry is represented as an offset of the density function $f_Z(\ln(\tau[s]))$. The peaks of the density functions of multi-Gaussian models (g_1 and g_2) are in between the peaks of the density functions of non-multi-Gaussian models and have a more symmetric shape. More slowest particles can be found in the non-multi-Gaussian models with a positive asymmetry (v_{11} and v_{21}) than in the multi-Gaussian models because the particles have a low probability to reach a blob with large-K. In a random field with a positive Asymmetry, the neighboring point of a large-K value has a higher probability of being a large-K value than a low-K value. Therefore, when particles travel in a blob with large-K, they can travel fast over a certain distance. So, a density of fast particles (small travel time) on the left side of the density

function can be found. In contrast, more fast particles can be found in the models with negative asymmetry (v_{12} and v_{22}) than the multi-Gaussian models because the particles have a low probability to touch a small value blob. However, when a particle is injected or travels into a low-K blob, it is "trapped" in the low-K zone, which leads to a large value of travel time.

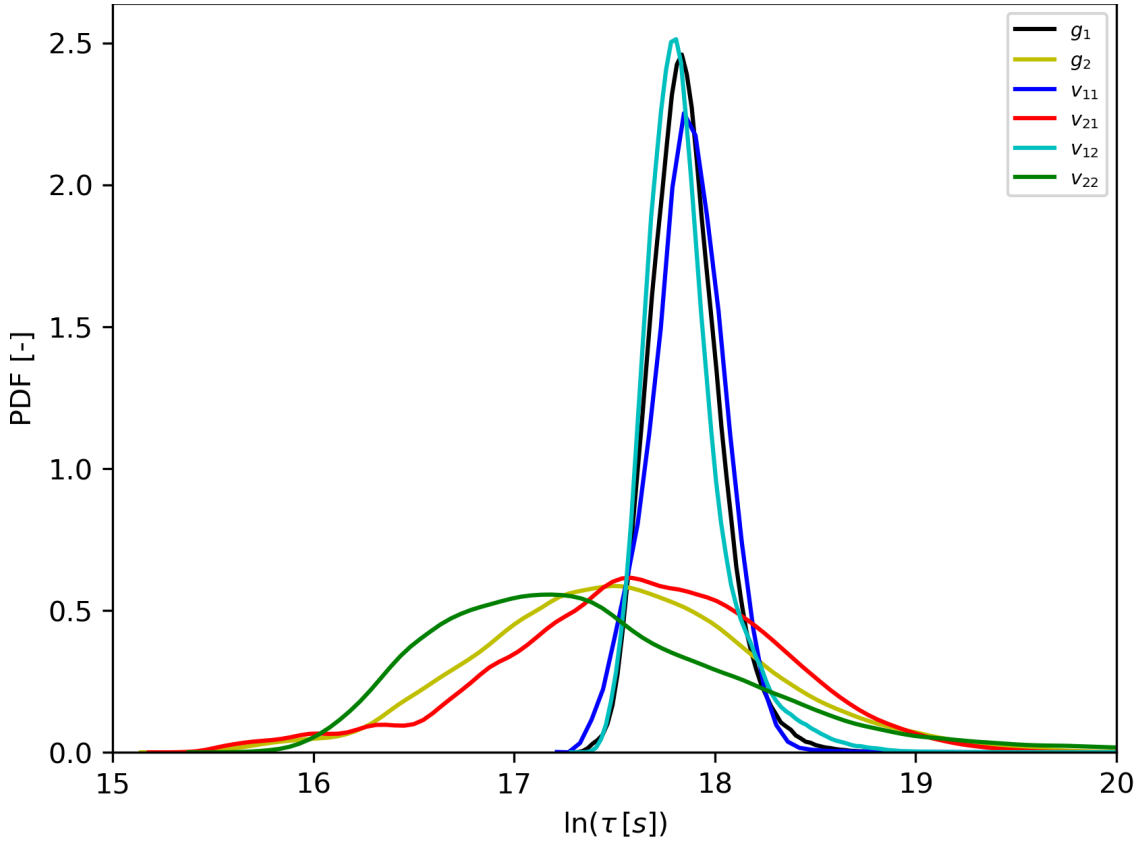

Figure 4.2: Distribution functions of the log-travel time $\tau[s]$ of multi-Gaussian (g_1-g_2) and non-multi-Gaussian (positive asymmetry: v_{11} and v_{21}, negative asymmetry: v_{12} and v_{22}), fields with a mild $\sigma^2(\ln(K))$ (g_1, v_{11} and v_{22}) and a large $\sigma^2(\ln(K))$ (g_2, v_{21} and v_{22}) at $x_2 = 150\,[m]$.

Other ensemble measures, e.g., the longitudinal dispersivity and velocity, are influenced by the variation in the particle travel time. Figure 4.3 shows the longitudinal ensemble dispersivity $\frac{D_\ell}{U_\ell}$ and velocity U_ℓ of various models. The influence of the included types of asymmetry does exist in all models. Although the absolute variation between a multi-Gaussian and non-multi-Gaussian model with a small $\sigma^2(\ln(K))$ (g_1, v_{11} and v_{12}) is less than the models with a large $\sigma^2(\ln(K))$ (g_2, v_{21} and v_{22}). When the difference between the multi-Gaussian and the non-multi-Gaussian model is large enough, if a multi-Gaussian model is used to describe a non-multi-Gaussian K-field, the uncertainty, which is caused

by the incorrect selection of the model, can not be neglected for a dataset with a large $\sigma^2(\ln(K))$.

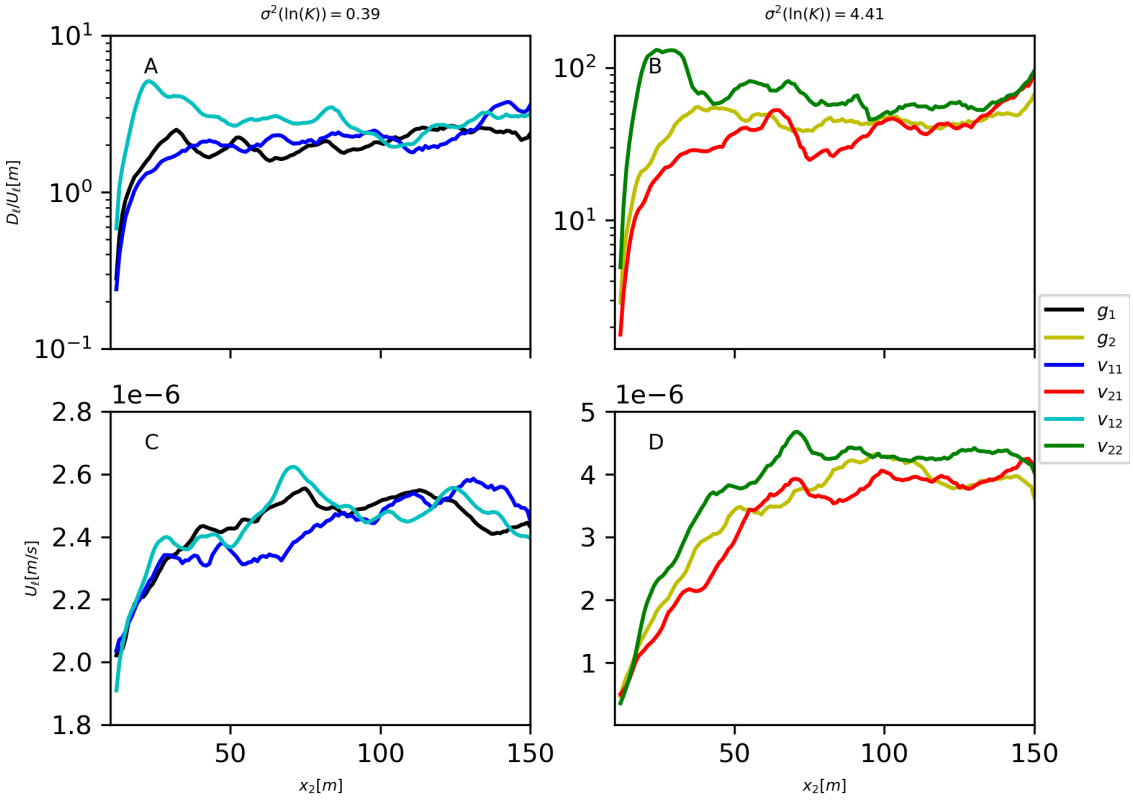

Figure 4.3: The ensemble longitudinal dispersivity of A) models with the marginal distribution M_1 (g_1, v_{11} and v_{12}); B) models with the marginal distribution M_2 (g_2, v_{21} and v_{22}); and the longitudinal velocity of C) models with the marginal distribution M_1 (g_1, v_{11} and v_{12}); D) models with the marginal distribution M_2 (g_2, v_{21} and v_{22}).

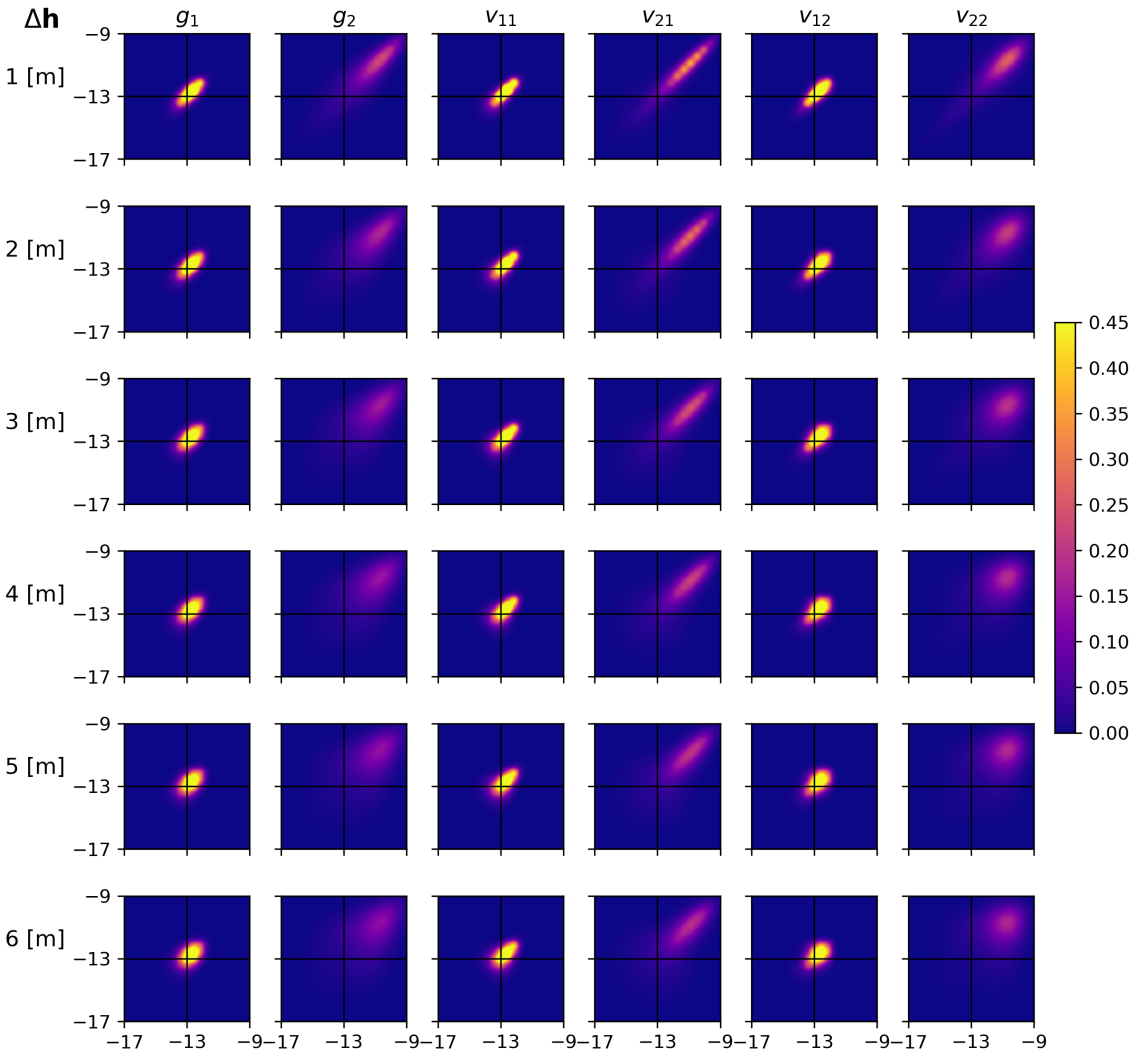

Figure 4.4: Joint probability density functions of the logarithm of the longitudinal particle velocity of multi-Gaussian (g_1-g_2) and non-multi-Gaussian (positive asymmetry: v_{11} and v_{21}; negative asymmetry: v_{12} and v_{22};) models with a mild $\sigma^2(\ln(K)) = 0.39$ (g_1, v_{11} and v_{22}) and a large $\sigma^2(\ln(K)) = 4.41$ (g_2, v_{21} and v_{22}) at different separation distances($1\,[m] \sim 6\,[m]$ from top row to the bottom row).

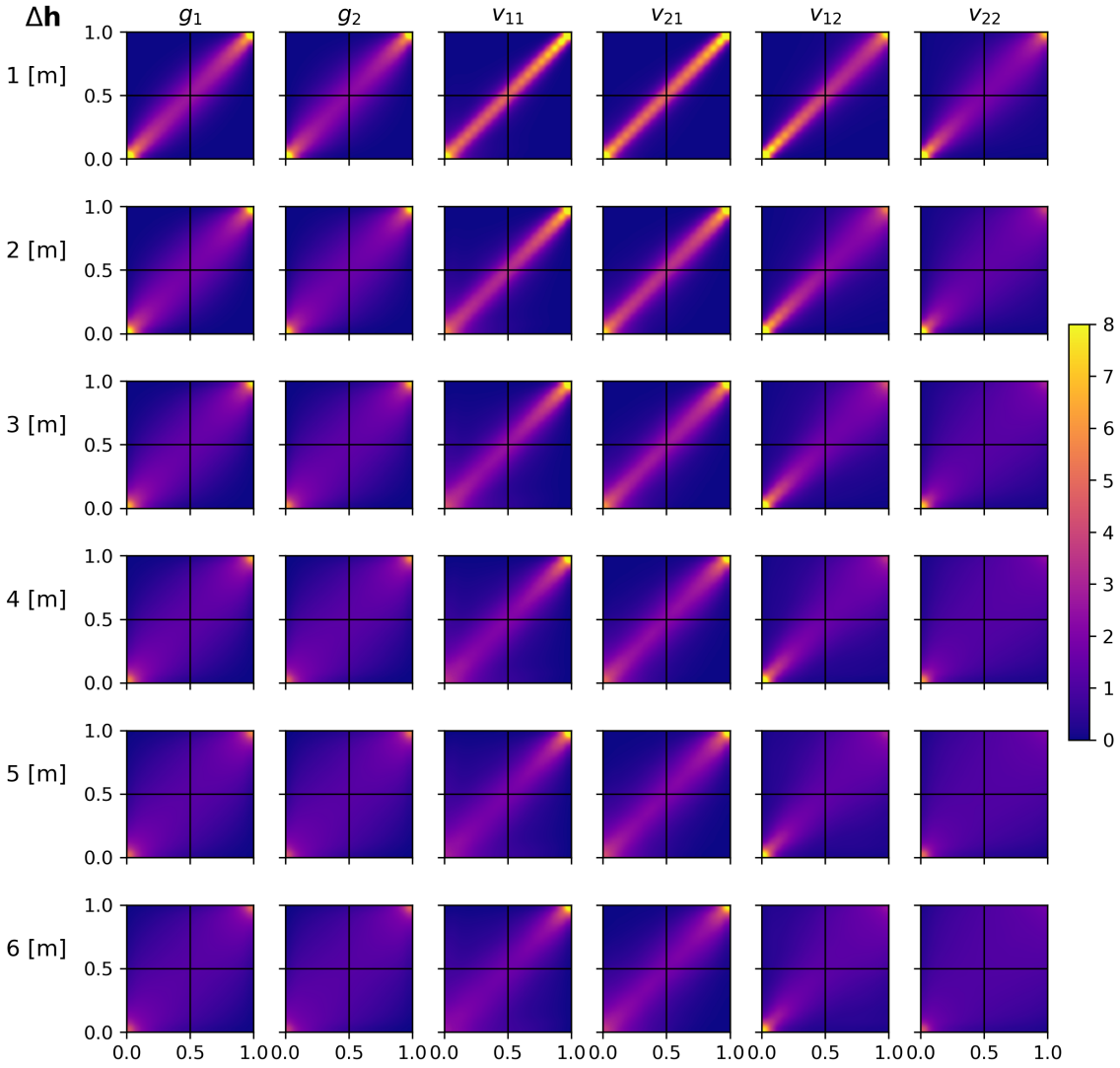

Figure 4.5: Bivariate copula densities of the longitudinal particle velocity of multi-Gaussian (g_1-g_2) and non-multi-Gaussian (positive asymmetry: v_{11} and v_{21}; negative asymmetry: v_{12} and v_{22};) models with a mild $\sigma^2(\ln(K)) = 0.39$ (g_1, v_{11} and v_{22}) and a large $\sigma^2(\ln(K)) = 4.41$ (g_2, v_{21} and v_{22}) in copula space at different separation distances($1\,[m] \sim 6\,[m]$ from top row to the bottom row).

The distribution of the longitudinal particle velocity plays the most important role in the particle-tracking simulation. Therefore, understanding the correlation of the longitudinal particle velocity on two observation planes with a certain separation distance can support a detailed understanding of the influence of the geostatistical models of the K-field and further modeling of the particle velocity. To analyze the correlation of the longitudinal particle velocity, the ensemble joint PDFs in the logarithm space and bivariate copula

densities with different separation distances ($1\,[m] \sim 6\,[m]$) are plotted in Figure 4.4 and Figure 4.5. A more spread-out bivariate distribution and copula density can be found for large separation distances than small separation distances, which indicates a stronger correlation between the longitudinal velocity of neighboring particles than the particles far away from each other.

In Figure 4.4, the joint density function is mainly influenced by the marginal distribution. Therefore, the models with the same marginal distribution (Table 4.1) have similar bivariate densities between each other. This influence from the marginal distribution is reduced after transformation into the copula space (Figure 4.5), in which the difference between different models is mainly driven by the underlying degree of Gaussianity of K. The multi-Gaussian models (g_1 and g_2) have a similar symmetric bivariate pattern between the top right corner (large-large velocities) and the bottom left corner (small-small velocities). In contrast, the large-large velocities have a stronger correlation in the models with a positive asymmetry (v_{11} and g_{21}) and the small-small velocities have a stronger correlation in the models with a negative asymmetry (v_{12} and g_{22}). Therefore, in case the correlation structure of the longitudinal particle velocity is mainly controlled by the degree of Gaussianity, i.e., an asymmetry away from zero and a large $\sigma^2(\ln(K))$, the copula density of the longitudinal particle velocity is a better choice to model the correlation structure of longitudinal particle velocity than the bivariate density, even in the logarithm space.

The potential use of copula densities such as Figure 4.5 is to simulate the particle velocity on the next observation plane when the particle velocity on the current observation plane is known. The transition of the longitudinal particle velocity between observation planes can be assumed as an n-step spatial Markov process (Le Borgne *et al.*, 2008). Then the longitudinal particle velocity on the following observation plane can be simulated according to the bivariate copula density and the corresponding particle velocity on the current observation plane.

Algorithm 5 shows an example method of a one-step Markov Chain simulation. The one-step empirical copula density (first row in Figure 4.5) is used as the prior information. The particle velocity of the current location $U_{\ell,i}$ is calculated from the conditional inverse copula $C_s^{-1}(u_2|u_1 = u(U_{\ell,i-1}))$ of the previous velocity $U_{\ell,i-1}$

$$\begin{aligned} C_s^{-1}(u_2) &= C_s^{-1}(u_2|u_1 = u(U_{\ell,i-1,j})) \\ U_{\ell,i,j} &= C_s^{-1}(u_2 = u^*) \end{aligned} \tag{4.9}$$

in which u^* is a uniformly distributed random number between 0 and 1.

Figure 4.6 shows the preliminary results of the copula densities based on the simulated particle velocities of different K models. The correlation structures of the large-large

Algorithm 5: One-step simulation of particle velocity using copula density.

Input: Empirical bivariate copula density $c_s(u_1, u_2)$ according to the particle tracking simulation
Result: Simulated particle velocities U_ℓ in copula space $u(U_\ell)$
$u(U_{\ell,1}) \leftarrow$ **Rand**$[0,1]$ initial velocities as a uniform distribution between 0 and 1;
loop along the main flow direction ;
for *i in* $[2, x_2]$ **do**
  loop along each particle ;
  for *j in* $[1, N_p]$ **do**
    $u(U_{\ell,i,j}) \leftarrow$ **Rand**$[0,1]$ with weights $c_s(u_2|u_1 = u(U_{\ell,i-1,j}))$;
  end
end

velocity and small-small velocity of different models can be caught in the simulated copula density. For further comparison, the χ^2 test is performed on the simulated copula density and the empirical copula density. The results in Figure 4.7 show that the test statistical values $\ln(\chi^2)$ of all models are under the critical line, which means the correlation structure of the longitudinal particle velocity at a large separation distance can be simulated using the one-step copula density.

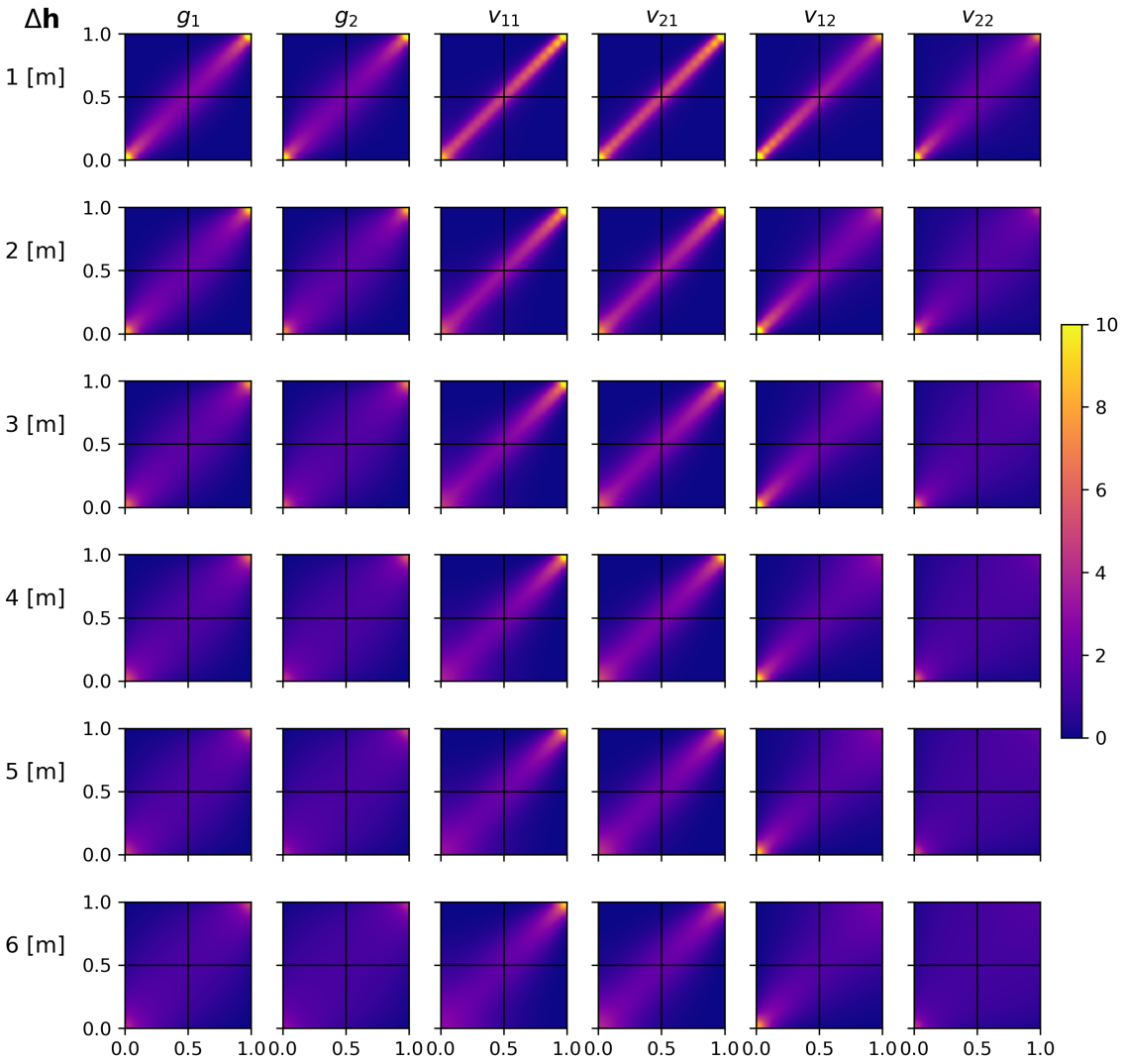

Figure 4.6: Bivariate copula densities of the longitudinal particle velocity of multi-Gaussian (g_1-g_2) and non-multi-Gaussian (positive asymmetry: v_{11} and v_{21}; negative asymmetry: v_{12} and v_{22};) models with a mild $\sigma^2(\ln(K)) = 0.39$ (g_1, v_{11} and v_{22}) and a large $\sigma^2(\ln(K)) = 4.41$ (g_2, v_{21} and v_{22}) with the one-step model at different separation distances($1\,[m] \sim 6\,[m]$ from top row to the bottom row).

Figure 4.8 shows the deviations between the simulated copula densities (Figure 4.6) and the reference copula densities (Figure 4.5). Most of the non-multi-Gaussian models (v_{11}, v_{12} and v_{21}) have a larger deviation than the multi-Gaussian models (g_1 and g_2). Possible sources of this deviation are: 1) the initial particle velocity is assumed as a uniform distribution between 0 and 1. 2) The empirical histograms with 20 bins are employed as the approximation of the copula density. A theoretical copula model could be used in the future to improve the estimation of the conditional copula density. 3) the directional

copula density could be used in the future work, which means $u_i \rightarrow u_j \neq u_j \rightarrow u_i$.

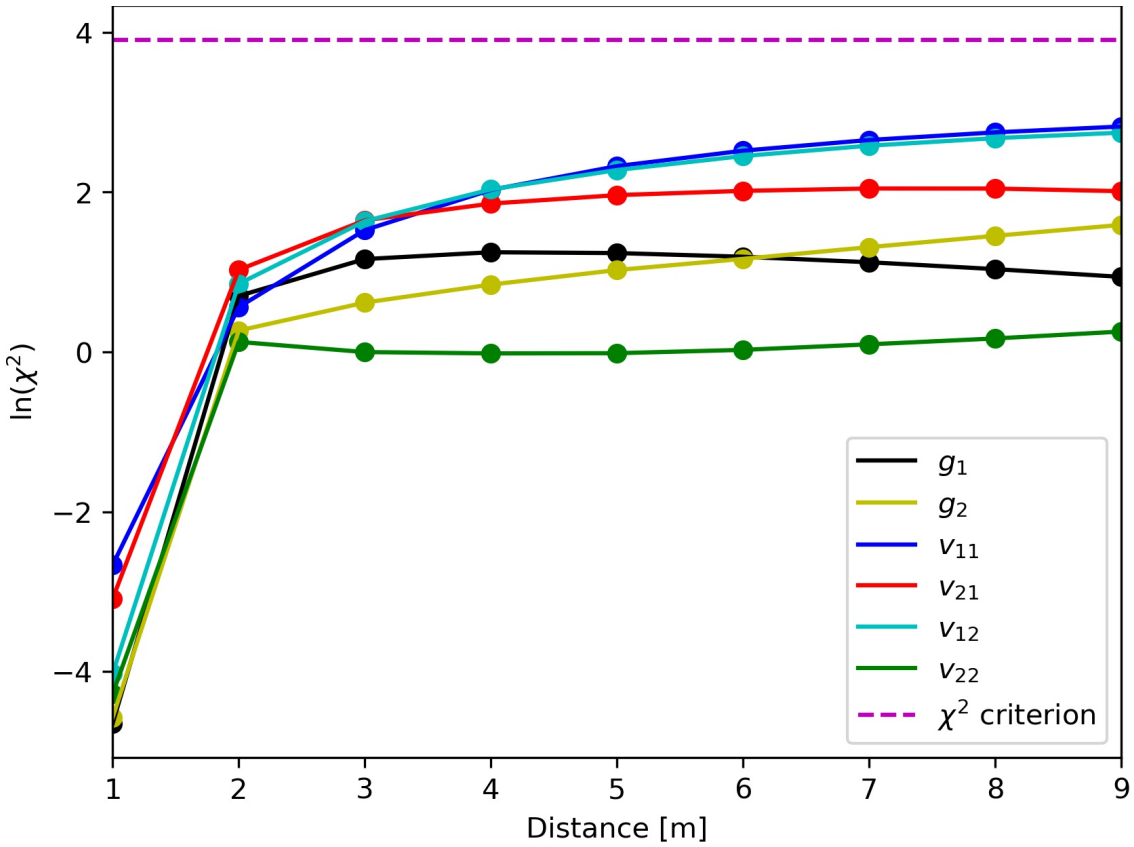

Figure 4.7: χ^2-test with a significance level 0.05 of the empirical bivariate copula densities and the simulated bivariate copula density with the one-step model of the longitudinal particle velocity.

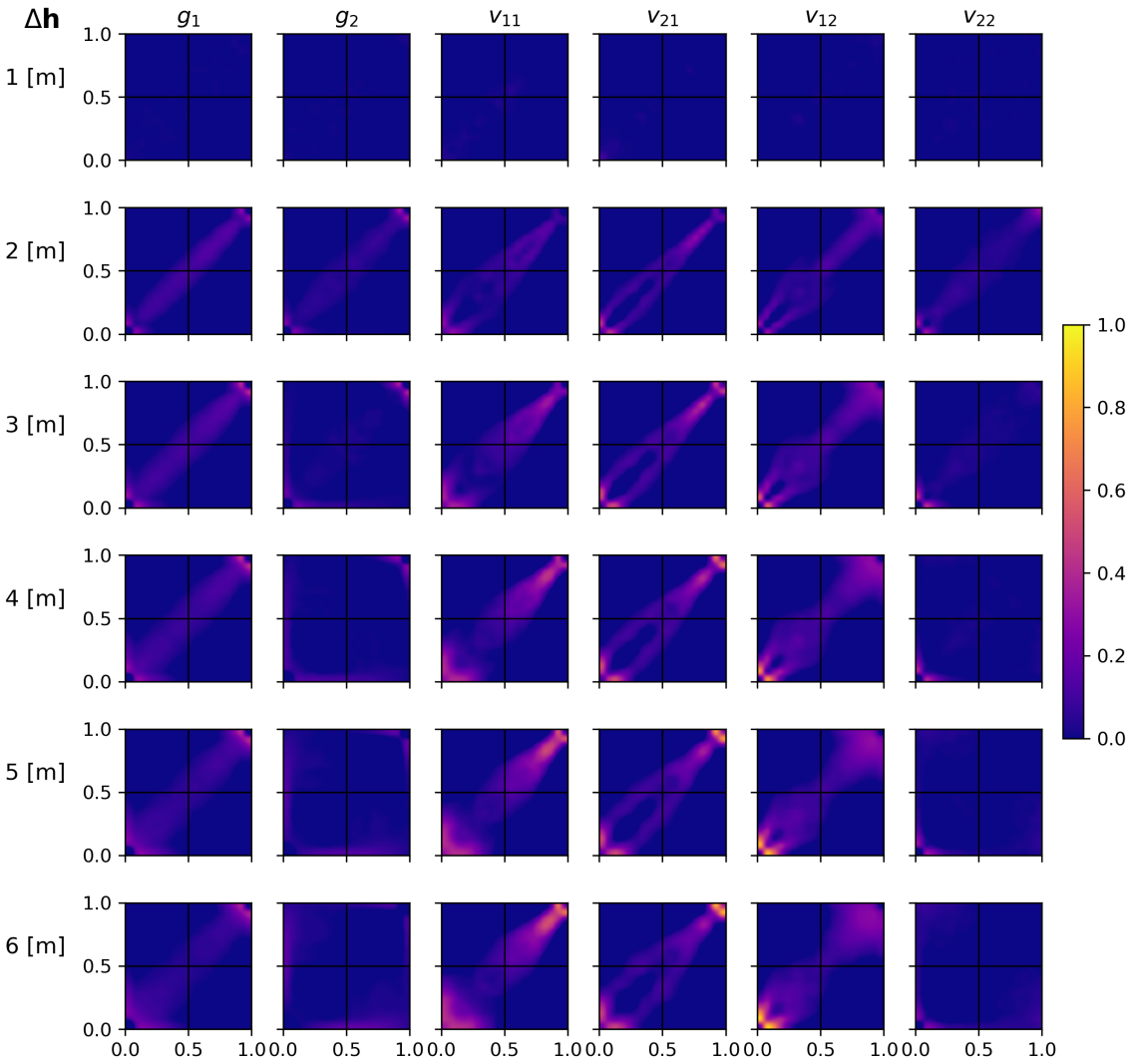

Figure 4.8: Deviations between the simulated bivariate copula densities with one-step simulation and the reference empirical bivariate copula densities of the longitudinal particle velocity of multi-Gaussian (g_1-g_2) and non-multi-Gaussian (positive asymmetry: v_{11} and v_{21}; negative asymmetry: v_{12} and v_{22};) models with a mild $\sigma^2(\ln(K)) = 0.39$ (g_1, v_{11} and v_{22}) and a large $\sigma^2(\ln(K)) = 4.41$ (g_2, v_{21} and v_{22}) with the one-step model at different separation distances($1\,[m] \sim 6\,[m]$ from top row to the bottom row).

4.3 Summary of this Chapter

In this chapter, the theoretical background of the particle-tracking random-walk method, which is used to evaluate the solute transport in this thesis, has been presented in Sec-

tion 4.1. The results of the test scenarios in Section 4.2 show that this method can catch the ensemble properties of the particles, which are influenced by geostatistical models of the K-field with different marginal distributions and degree of Gaussianity. The non-multi-Gaussian spatial dependence leads to a deviation of the ensemble measures of the multi-Gaussian models. This deviation is proportional to the $\sigma^2(\ln(K))$. Therefore, for the models with a large $\sigma^2(\ln(K))$, the deviation caused by a non-multi-Gaussian spatial dependence can not be neglected.

Furthermore, the bivariate copula density of the longitudinal particle velocities provides information to model the behavior of particle velocities under different K models. A theoretical model between the Gaussianity and the corresponding bivariate copula density could be further developed in future work.

Chapter 5

Application to the MADE Site: Data Description and (Geo-) Statistical Evaluation

The content in this chapter contains materials published in "Xiao, B., Haslauer, C., and Bohling, G. (2019). Comparison of multivariate spatial dependence structures of DPIL and flowmeter hydraulic conductivity data sets at the MADE site. Water (MDPI), 11(7), 1420".

Author	**Author position**	**Scientific ideas %**	**Data generation %**	**Analysis & interpretation %**	**Paper writing %**
Bo Xiao	1	70	80	60	75
Claus Haslauer	2	20	20	30	20
Geoffrey Bohling	3	10	0	10	5
Titel of paper:	Comparison of multivariate spatial dependence structures of DPIL and flowmeter hydraulic conductivity data sets at the MADE site.				
Status in publication process:	Published.				

In this chapter, two non-colocated K datasets of flowmeter measurements and direct-push injection-logging (DPIL) at the MADE site are compared using various (geo-) statistical measures. After a short introduction of the two datasets at the MADE site in Section 5.1, the univariate statistical measures of the two datasets and their distributions in the vertical and horizontal directions are compared in Section 5.2. In Section 5.3, the copula-based geostatistical measures of the two datasets are compared. Then in Section 5.3.2 the results of the copula-based parameter estimation using the maximum likelihood method in Sections 2.4.2 and 2.4.3 are presented. A short conclusion and summary is given in

Section 5.4 .

5.1 The MAcroDispersion Experimental Data Set

Figure 5.1: Locations of the DPIL (blue cross sign) and flowmeter (brown plus sign) boreholes at the MADE site.

Two K datasets are available at the MADE site. One dataset was measured using a borehole flowmeter (Boggs *et al.*, 1992) at 2611 observation points in 67 wells with a vertical spacing of 15 [cm], the layer hydraulic conductivity is calculated from the discharge from each horizontal layer entering the well. Another dataset was obtained using the high-resolution direct-push injection-logger (DPIL) at 31,123 observation points in 58 vertical profiles with a vertical spacing of 1.5 [cm]. The DPIL ratio (injection rate/pressure ratio) is taken as a measure of the relative hydraulic conductivity. This DPIL ratio has been calibrated using the colocated direct push permeameter (Liu *et al.*, 2009; Bohling *et al.*, 2012) and later reassessed after considering the measured noise and the upper limit of the DPIL ratio (Bohling *et al.*, 2016). Figure 5.1 shows the measurement locations of the flowmeter and the DPIL dataset using a modified MADE-2 coordinate system (following Boggs *et al.*, 1993). The DPIL dataset has many measurements in the vertical direction and the flowmeter profiles provide better lateral coverage of the site than the DPIL profiles (Figure 5.1).

5.2 Statistical Evaluation of the MADE Data Sets

In this Section, univariate statistical measures (Section 5.2.1) of the DPIL and flowmeter datasets and their distributions in space (Section 5.2.2) are presented and compared.

5.2.1 Marginal Distribution and Basic Statistics

Table 5.1: Univariate statistical measures of the flowmeter, DPIL dataset at the MADE site, and the Borden dataset.

	MADE flowmeter	MADE DPIL	Borden
Geometric mean [m/s]	4.29×10^{-5}	6.73×10^{-6}	9.54×10^{-5}
Arithmetic mean $[\ln(K[m/s])]$	-9.26	-10.06	-11.91
Min [m/s]	1.83×10^{-7}	1.46×10^{-9}	5.67×10^{-7}
Max [m/s]	1.45×10^{-2}	1.96×10^{-2}	3.29×10^{-4}
Variance $[\ln(K[m/s])]^2$	4.41	5.91	0.39
Skewness [m/s]	6.28	15.98	0.80
Kurtosis [m/s]	58.74	322.66	0.47

Table 5.1 lists the univariate statistical measures of the flowmeter and the DPIL datasets at the MADE site together with the Borden dataset (Sudicky, 1986) as a reference. Both the flowmeter and DPIL dataset show that the MADE site is more heterogeneous than the Borden site, e.g., a wide range of the K values (large maximal K and small minimal K), a large $\sigma^2(\ln(K))$, skewness and kurtosis. Within the two datasets at the MADE site, the DPIL dataset has a larger $\sigma^2(\ln(K))$ and one order of magnitude smaller geometric mean than the flowmeter dataset, although both datasets measure the same variable at the same site.

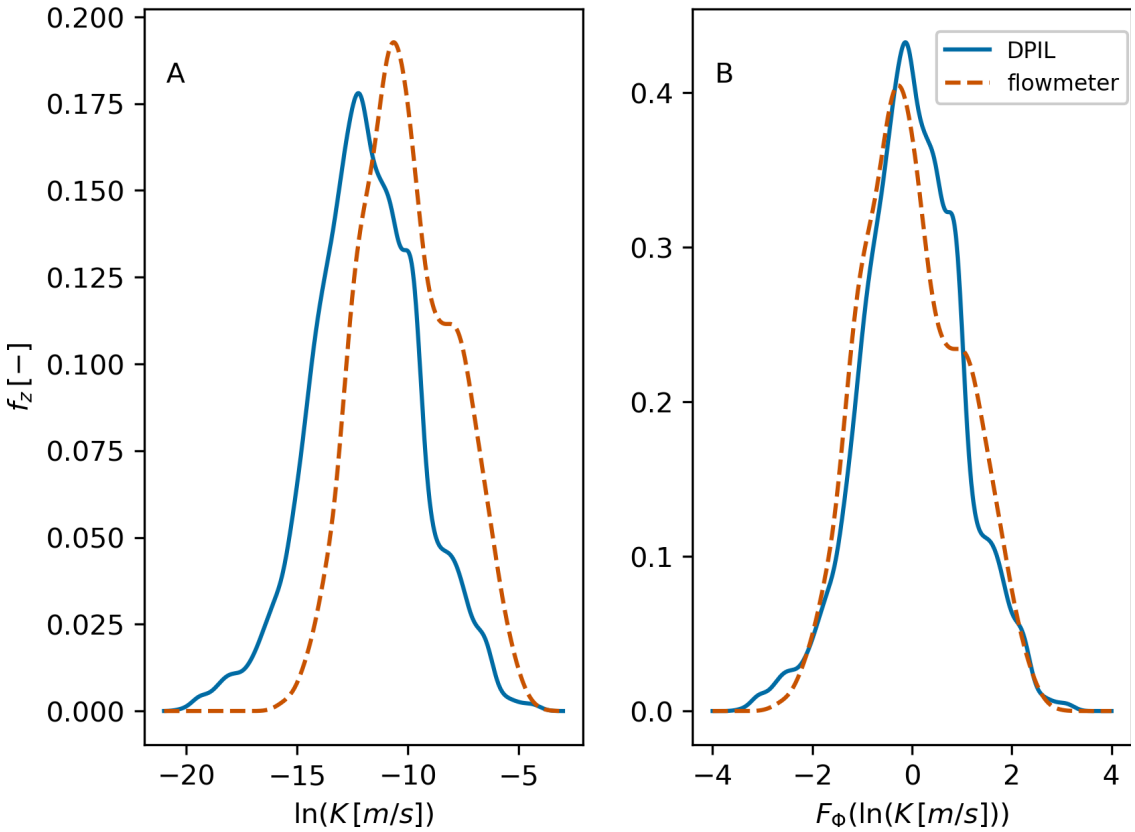

Figure 5.2: Probability density function of $\ln(K)$ of DPIL and flowmeter dataset. A) PDF of $\ln(K)$ and B) PDF of $\ln(K)$ after the standard transformation.

This difference between the two datasets is also visible in the plots of the probability density function of log-K (Figure 5.2A). Both datasets have probability density functions with similar shapes despite a shifted mean value. The DPIL dataset has a larger portion of lower K values, a smaller portion of larger K values, and a larger spread than the flowmeter dataset. This finding is in congruence with Bohling *et al.* (2016). Even after performing a standard normal transformation ($F_\Phi = \frac{z(\mathbf{x})-\mu}{\sigma}$) on both datasets, a difference of the density function can be found between $0.0\,[-]$ and $2.0\,[-]$ (Figure 5.2B). An obvious question due

to this difference is whether the two datasets measure the same K-field or not. Therefore, the key point in the following sections is about defining the differences and similarities between two datasets using different (geo-)statistical measures.

5.2.2 Spatial Distribution of K Observations

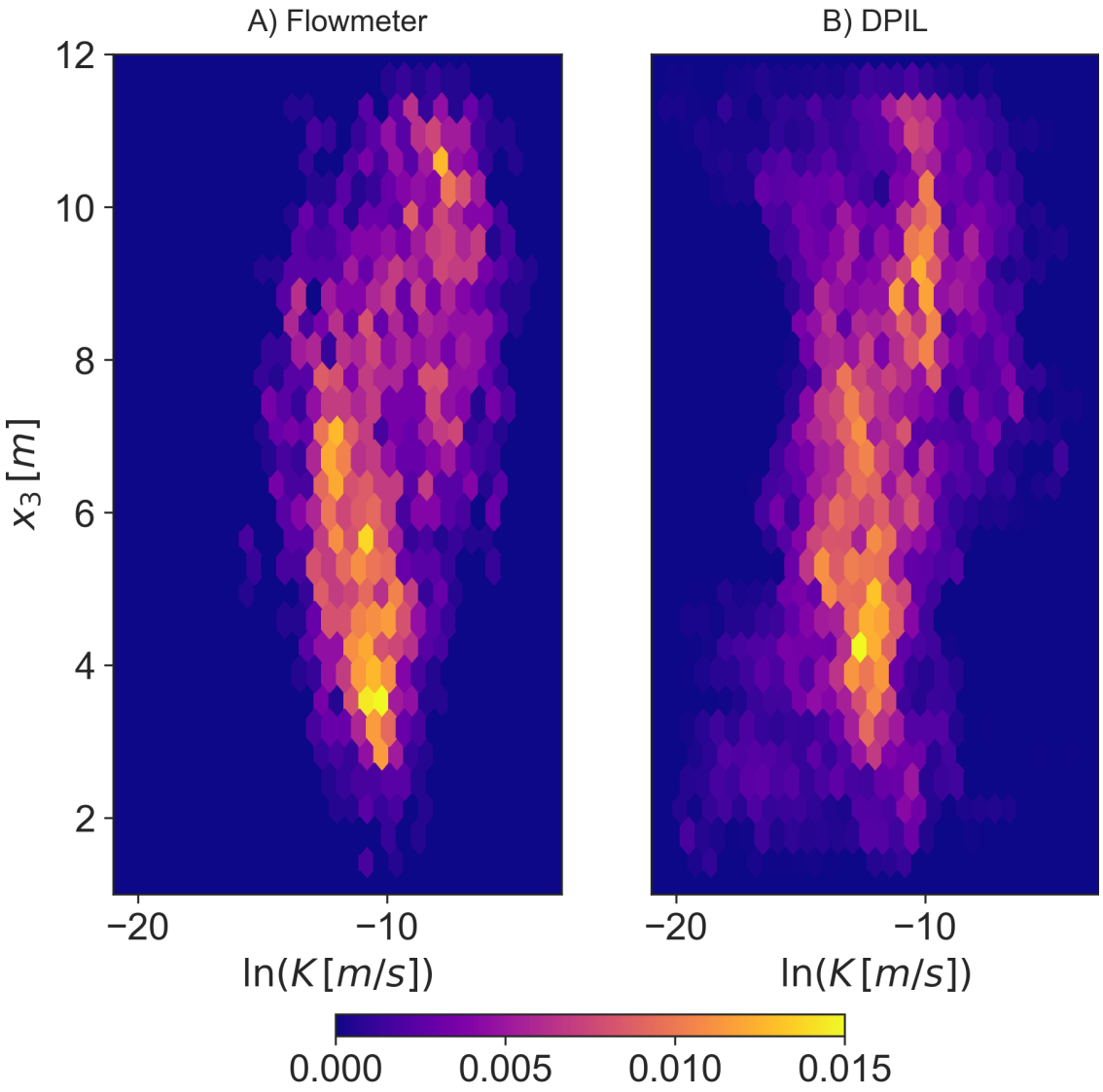

Figure 5.3: Bivariate density functions of elevation x_3 vs. $\ln(K)$ for the A) flowmeter dataset B) DPIL dataset. Elevation given in meters above datum; Datum is 14.1 m below reference.

The shifted mean value found in the marginal distributions (Figure 5.2) can be also seen on the plot of two-dimensional histograms of $\ln(K)$ versus vertical elevation x_3 (Figure 5.3). The DPIL dataset has smaller K values in general than the flowmeter dataset. Both plots of the two-dimensional histogram show a fairly distinct shift at about 8 [m] above the reference elevation (see Bohling *et al.* (2016) for a definition of the coordinate system), with generally large K values above this elevation and generally small K values below. Below 8 [m] elevation, the flowmeter data seem to indicate a slightly increasing trend of K with increasing depth, a trend which might also be reflected in the DPIL data if one considers the higher density region of the data, discounting the relatively small proportion of very low DPIL K values at elevations $x_3 < 5\,[m]$.

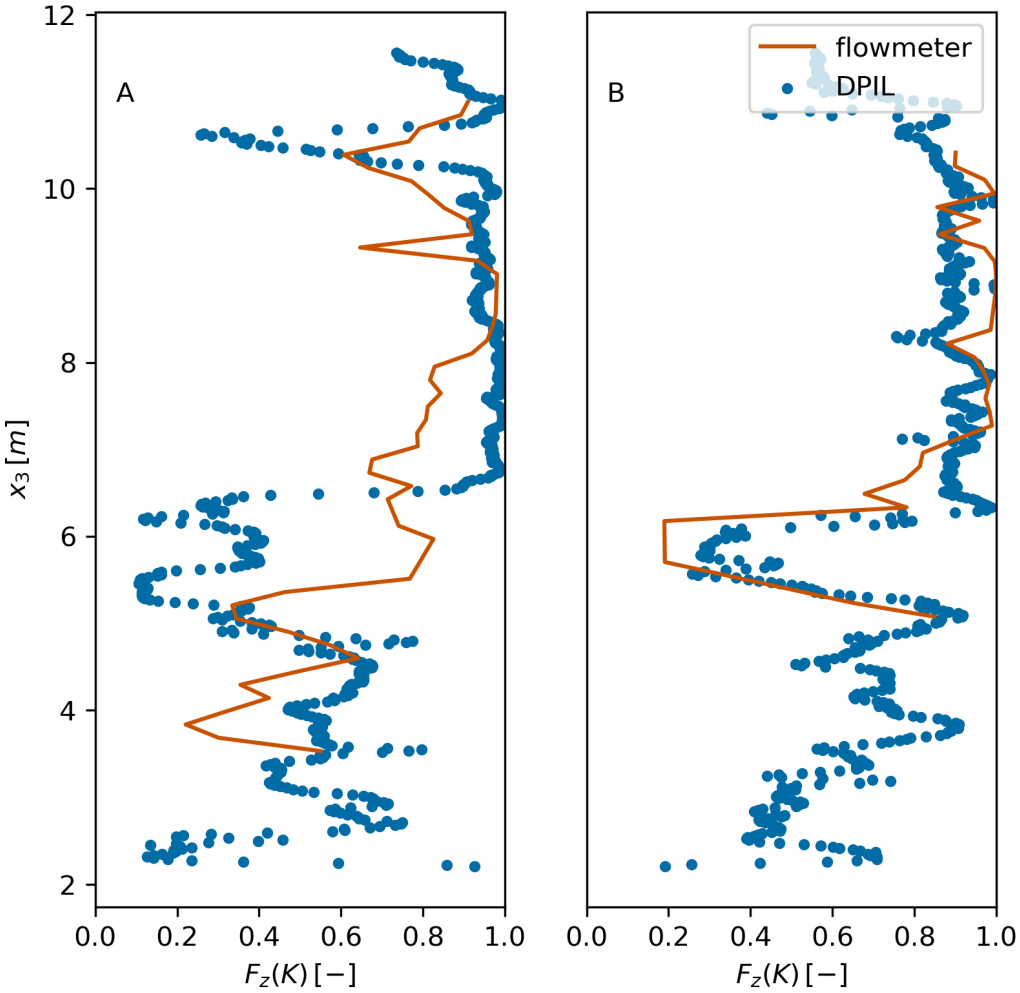

Figure 5.4: Comparison between the DPIL and flowmeter datasets from the nearest boreholes: A) flowmeter:$x_1 = 44.82\,[m], x_2 = 87.82\,[m]$; DPIL:$x_1 = 44.66\,[m], x_2 = 88.27\,[m]$; $h = 0.47\,[m]$; B) flowmeter:$x_1 = 48.76\,[m], x_2 = 90.47\,[m]$; DPIL:$x_1 = 48.46\,[m], x_2 = 90.76\,[m]$; $h = 0.41\,[m]$.

To reduce the influence of the marginal distribution, the two datasets are compared using the distribution function $F_Z(K)$ and accounting for the horizontal separation distance h. If both datasets measure the same K-field, a similarity would be found between the vertical K profiles of the two datasets when measured close to each other, although the flowmeter dataset measured a shorter interval in the vertical direction than the DPIL dataset, Figure 5.4 shows two comparisons between the K profiles with a horizontal separation distance less than 0.5 [m] (Figure 5.4A: $h = 0.47\,[m]$; B: $h = 0.41\,[m]$). A high similarity between the two datasets can be found in both plots, although a difference between the two datasets can be found in Figure 5.4A around $5 \sim 6\,[m]$.

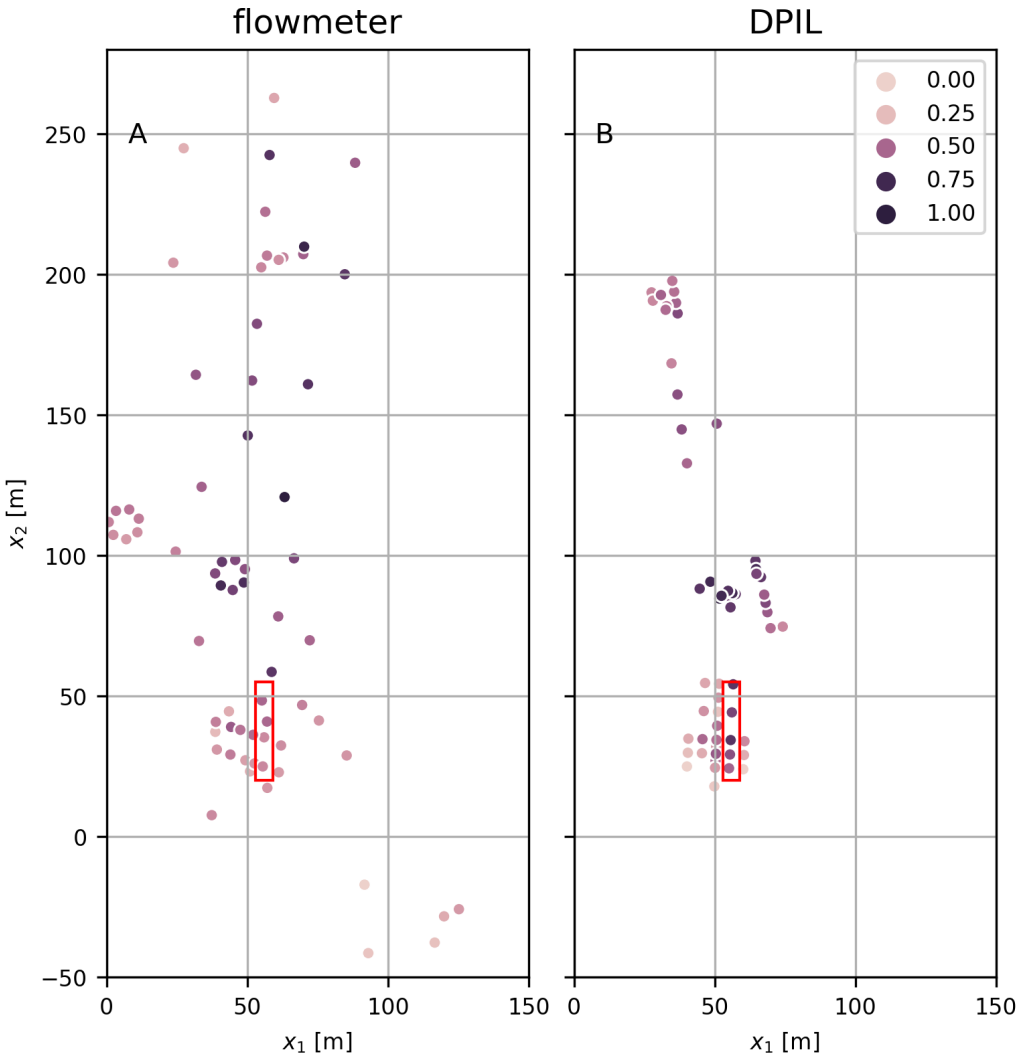

Figure 5.5: Vertical averaging of the distribution functions of A) flowmeter and B) DPIL datasets in the horizontal space. The red box indicates the locations of outliers.

Figure 5.5 shows the vertical averaging of the distribution function of the two datasets. In general, small values can be found on the bottom of both plots ($x_2 < 60\,[m]$) and large values can be found on the middle of both plots ($60\,[m] < x_2 < 200\,[m]$). Some outlier points can be found at the locations in the red boxes. Figure 5.6 shows the vertical profiles of these outliers. Observations of the DPIL dataset have larger values of $F_Z(K)$ than the flowmeter dataset, especially at the top and at the bottom. The reason for this difference is unknown but could be an interesting point for future fieldwork.

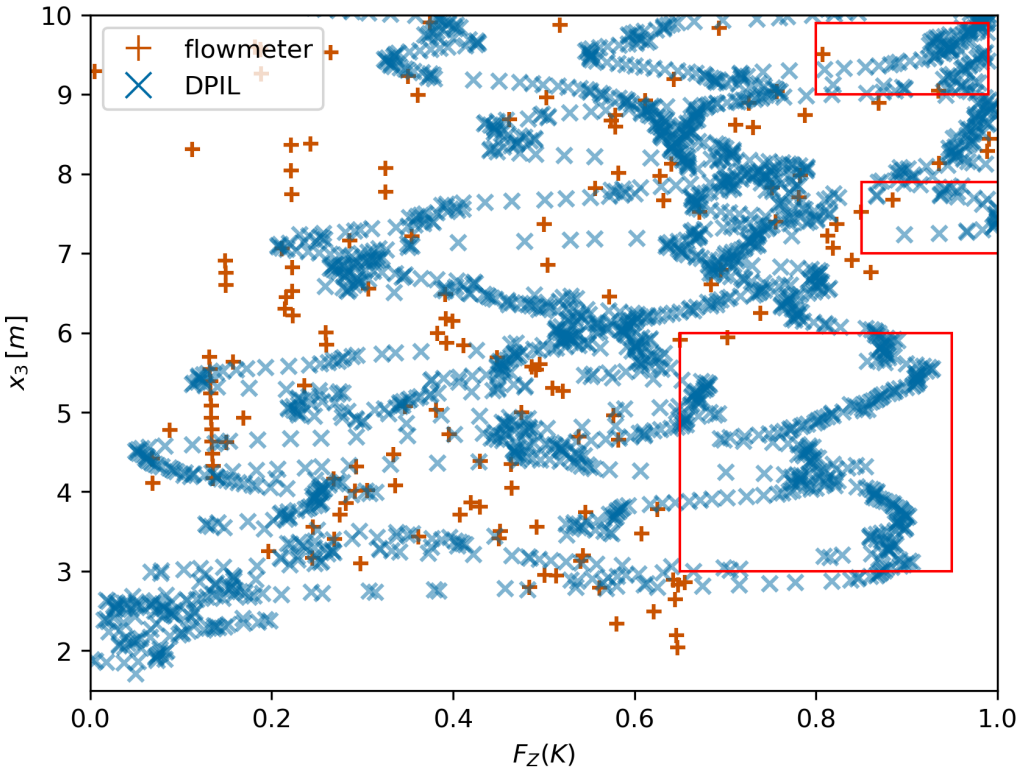

Figure 5.6: Vertical profile of two datasets on the locations with outliers in Figurer 5.5. The red boxes are the area with a larger DPIL profile than the flowmeter profile.

5.3 Geostatistical Evaluation of the MADE Data Set

5.3.1 Empirical Spatial Dependence

In this section, three empirical measures of spatial dependence, standardized variogram $\gamma(\mathbf{h})$, rank correlation $\rho_s(\mathbf{h})$, and asymmetry $A(\mathbf{h})$ (Figure 5.7), are analyzed and compared between the flowmeter and DPIL datasets. Additionally, empirical bivariate copula

densities are compared.

The spatial dependence structures have been analyzed in the vertical and horizontal directions and in an equivalent isotropic coordinate system obtained by geometrically stretching the vertical axis. The directional analysis assumes azimuthal isotropy (same range for all horizontal directions), with an elliptical representation of anisotropy in the vertical plane (shortest range in the vertical direction, longest in the horizontal). Our computations of the empirical measures of spatial dependence (either variogram or copula-based measures) have used the same data pair selection criteria as Rehfeldt *et al.* (1992). In the vertical direction, the tolerance angle is 0.1° and the tolerance bandwidth is $0.1\,[m]$ so that only within-borehole data pairs (no cross-borehole pairs) contribute to the vertical measure. For the horizontal measure, a vertical tolerance angle of 5° and a vertical bandwidth of $0.16\,[m]$ are used to limit the vertical separation between selected data pairs. After calculating the empirical rank correlation in the vertical and horizontal directions, the vertical to horizontal anisotropy ratio was estimated from the ranges. The data were then cast into an equivalent isotropic coordinate system by stretching the vertical axis by the estimated anisotropy ratio. In addition to the previously observed similarities (Bohling *et al.*, 2012), a systematic deviation of the asymmetry from zero was detected, indicating quantifiable deviance from Gaussian dependence. This will be discussed in the following part.

Pairwise Spatial Dependence The vertical normalized variogram γ, rank correlation ρ_s and asymmetry A for the two datasets are very similar (Figure 5.7, left column) and the equivalent isotropic versions of the variograms and rank correlations are also similar (Figure 5.7, right column). The DPIL dataset exhibits smoother results than the flowmeter dataset, as it is influenced by the finer vertical sample spacing of the DPIL dataset. A larger difference can be found in the horizontal direction (Figure 5.7, central panel). This difference in the horizontal direction may be caused by the difference between the lateral distributions of the flowmeter and DPIL profiles, which can be expressed as the difference in the number of data pairs (bottom row on Figure 5.7) or as the difference in the properties of the sampled area (Figure 5.5). The rank correlations between horizontal lags of 2.5 [m] and 17.5 [m] are closer to each other than the normalized variograms, most likely because of the smaller impact of outliers in rank space.

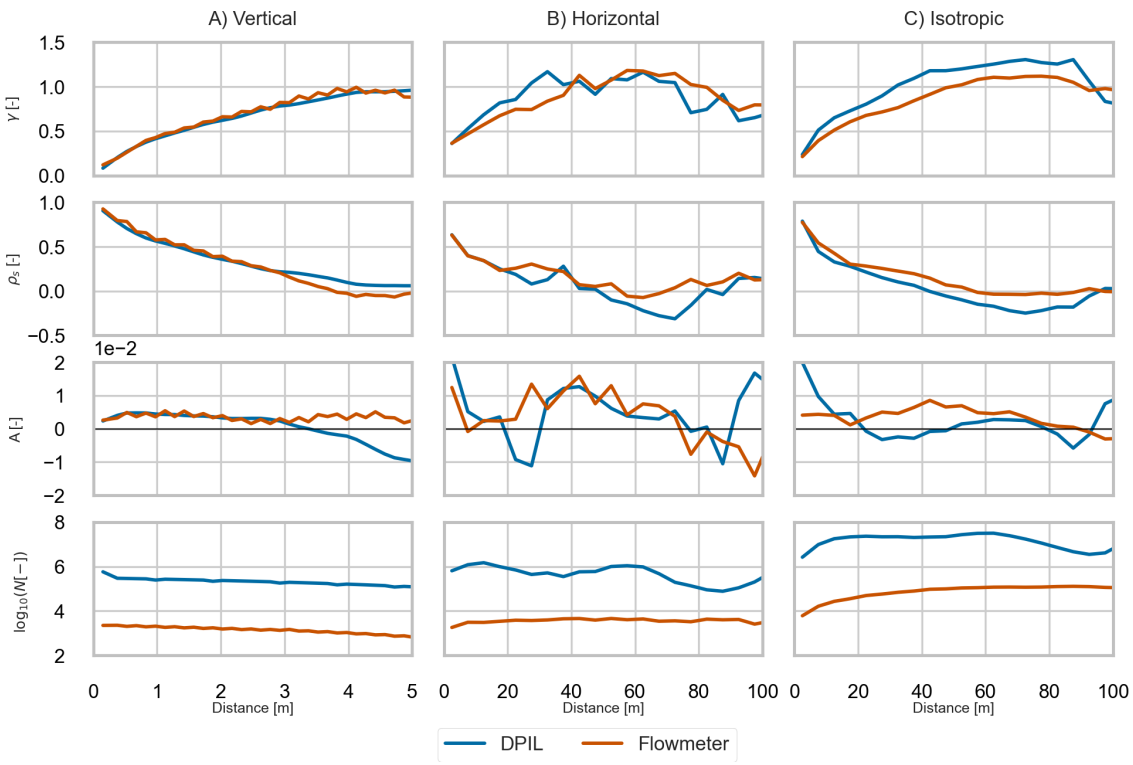

Figure 5.7: Bivariate measures in the vertical direction (panel **A**), the horizontal direction (panel **B**) direction, and the isotropic case (panel **C**). From top to bottom are: normalized variogram, copula-based rank correlation ρ_s, asymmetry A, number of data pairs N.

Asymmetry The copula asymmetry A (third row in Figure 5.7) differs consistently from zero for both datasets. A bootstrap analysis, involving a random selection of 70% of the data in each realization, indicates that the deviation from zero is consistent for each dataset (Figure 5.8), providing compelling evidence that the MADE site K-field exhibits non-multi-Gaussian spatial dependence. The narrower uncertainty envelope for the DPIL results is due to the larger number of DPIL data.

In the vertical direction, the values of asymmetry for the two datasets are essentially identical up to a lag of 3 [m]. For larger separation distances, the value of asymmetry of the DPIL data decreases, reaches zero at 4 [m], and becomes negative beyond 4 [m]. By contrast, the asymmetry remains positive for the flowmeter data. The observed behavior of asymmetry indicates a stronger dependence among lower K values for the DPIL dataset and among larger K values for the flowmeter dataset at large lag distances. This difference might be caused by the underestimation of high K values by the DPIL and overestimation of low K values by the flowmeter (Bohling *et al.*, 2012). Another possible reason is the difference in the distribution of the observations in the vertical direction, although

the DPIL data have been trimmed to match the vertical extent of the flowmeter data sitewide (Bohling *et al.*, 2012), the vertical extents of the flowmeter and DPIL profiles still differ locally. This may lead to a difference in the density of higher and lower K values (Figure 5.3) in the two datasets, contributing to the differences in the values of asymmetry.

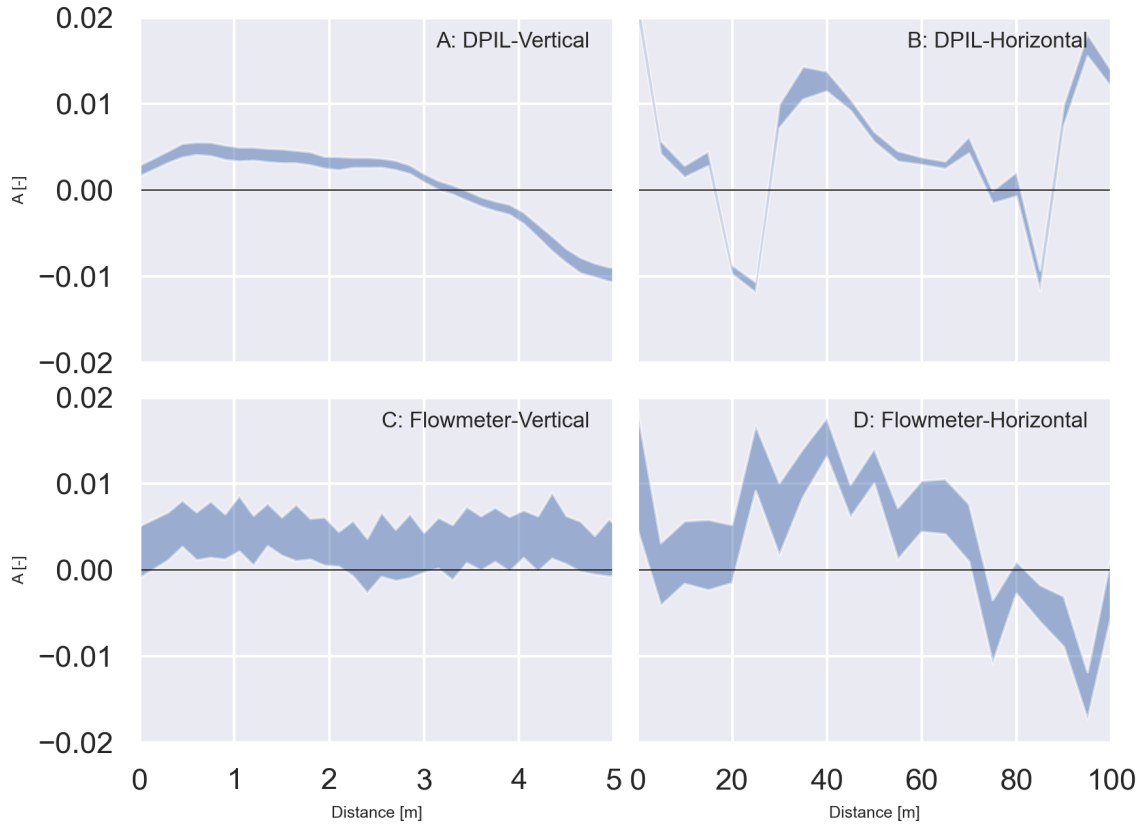

Figure 5.8: Value areas of the copula-based asymmetry (A) of bootstrapping tests with random selections of the 70% of the whole data set.

The bootstrap analysis (Figure 5.8) shows that the asymmetry A differs significantly from zero at most lags for both datasets, indicating that the MADE site K-field exhibits a nonsymmetric, and thus non-multi-Gaussian, spatial dependence structure.

Bivariate Empirical Copula Density To compare the bivariate spatial dependencies of the two datasets in more detail, the empirical bivariate copula densities of the first few lag distances have been plotted (Figure 5.9) and a χ^2 test with a significance level $\alpha = 0.05$ has been performed. The null hypothesis for this test is that the structure of the empirical bivariate copula densities of the two datasets for a certain lag distance are

the same. This hypothesis is rejected if the χ^2 statistic, representing the sum of squared differences between the two densities, is sufficiently large. In this case, the flowmeter and DPIL copula densities are deemed to be significantly different (the null hypothesis is rejected) only for the relatively large horizontal lag distances of 27.5 [m], 32.5 [m], and 37.5 [m] (Figure 5.10), where data availability is an issue (fourth row in Figure 5.7) and that have a lesser impact on solute transport than shorter lags. For all other lags, the copula densities are not deemed to be meaningfully different between the flowmeter and DPIL K datasets. This is further evidence that the flowmeter and DPIL datasets exhibit similar spatial dependency structures.

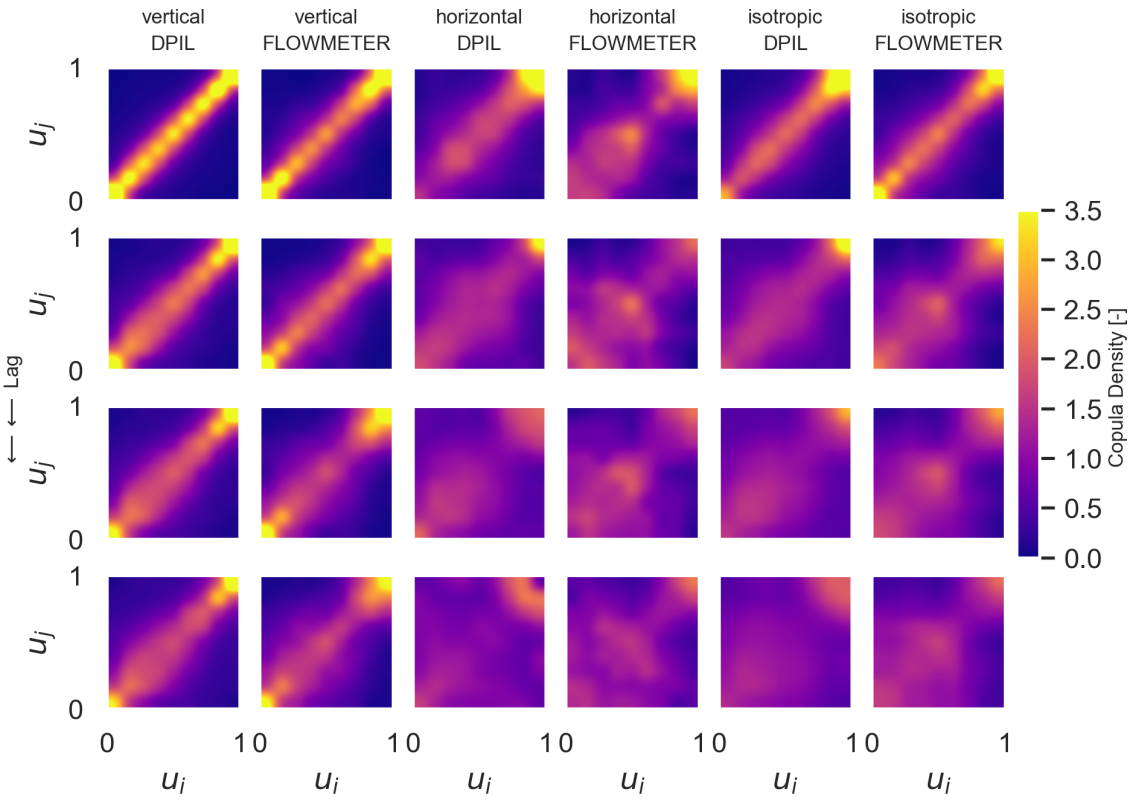

Figure 5.9: Empirical copula densities in the vertical direction (DPIL: 1st panel, flowmeter: 2nd panel), in the horizontal direction (DPIL: 3rd panel, flowmeter: 4th panel) and the isotropic case (DPIL: 5th panel, flowmeter: 6th panel). Distances from top to bottom are 0.15, 0.375, 0.525, 0.675 [m] in the vertical direction and 2.5, 7.5, 12.5, 17.5 [m] in the horizontal direction and the isotropic case.

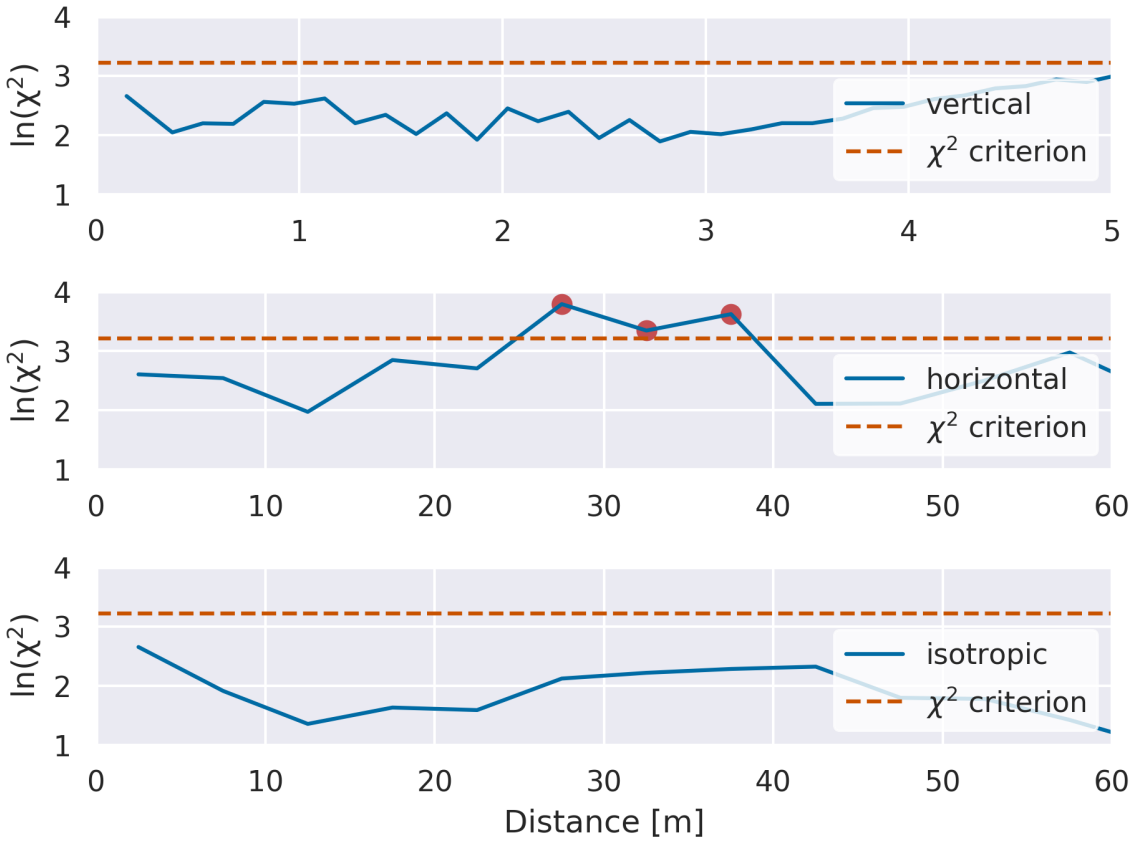

Figure 5.10: $\chi^2 - test$ with a significance level 0.05 in the vertical direction (**top**), the horizontal direction (**middle**), and the isotropic case (**bottom**). Distances with values larger than the χ^2 criterion (dash line) indicate dissimilar copula densities between two datasets (red points).

Isotropic measures of dependence were calculated based on geometrically transformed coordinates. After calculation of the directional measures, the anisotropy ratio of the two datasets can be calculated according to the results of ranges of the rank correlation in vertical (a_{Ver}) and in horizontal directions (a_{Hor}) (Table 5.2). The two datasets have similar values of a_{Hor} but different values of a_{Ver}, which leads to different values of the anisotropy ratio (a_{Hor}/a_{Ver}), 8.18 [-] for DPIL and 14.7 [-] for flowmeter. Then, vertical coordinates were stretched by using the anisotropy ratio to generate an isotropic dataset, such that the dataset after the transformation has identical vertical and horizontal ranges. The statistics for each isotropic case are calculated after that. In general, the two datasets exhibit similar behaviors of the normalized variogram, the rank correlation, and the asymmetry for the isotropic case.

Table 5.2: Distances where the rank correlation is zero (range) in the vertical (a_{Ver}) and horizontal (a_{Hor}) direction and the anisotropy ratio (a_{Hor}/a_{Ver}).

	$a_{Hor}\,[m]$	$a_{Ver}\,[m]$	$a_{Hor}/a_{Ver}\,[-]$
DPIL	48.3	5.91	8.18
Flowmeter	55.4	3.77	14.7

The foregoing analysis of bivariate measures has demonstrated a strong similarity between the two datasets for short separation distances, especially in the vertical direction, along with providing compelling evidence that both datasets exhibit significantly non-multi-Gaussian spatial dependence structures.

5.3.2 Results of the Copula Parameter Estimation

Results of the parameter estimation of DPIL and flowmeter dataset for spatial dependence using Gauss- and v-copula models are presented in this section. The analysis in this section is based on models with an optimized parameter vector $\boldsymbol{\theta}$ of n-point subsets of the data, providing a richer representation of the spatial dependence structure than the bivariate (two-point) measure (5.3.1).

As mentioned in Chapter 2, a maximal likelihood method was used to fit the theoretical copula models on randomly selected subsets. Further details of the fitting process follow:

Sample Spacing The vertical sample spacing of the DPIL measurements is one-tenth of the flowmeter measurements. To reduce the influence of this difference on the estimated parameters, the DPIL data were averaged with the same vertical spacing as the flowmeter data before fitting the copula models.

Dimension Parameter estimation was performed with varying numbers of data points included in each subset to provide an assessment of the effect of the subset size. The tested subset sizes were $n(S_w) \in (10, 12, 15)$. The estimated parameter values were found to be independent of $n(S_w)$ if $n(S_w) \geq 10$, a finding in accordance with Bárdossy and Li (2008). Consequently, results for the smallest sub-set size, $n(S_w) = 10$, are presented.

Parameter Vector $\boldsymbol{\theta}$ A covariance function of the Matérn type (Equation 2.14) with the range a and the smoothness parameter κ was used. For the case with a v-copula model, there were two additional parameters: m_c and k_c (Equation 2.32).

The estimated copula model parameters are shown in Tables 5.3 and 5.4 and the model-based rank correlations computed are compared with the empirical rank correlations in Figure 5.11. For the v-copula estimates, the parameter m_c is much smaller than 3.0 in all cases, indicating a non-multi-Gaussian type of dependence. To evaluate the performance of Gauss-copula and v-copula models, the Akaike information criterion ($AIC = -2\log L + 2n_{para}$ (Bozdogan, 1987), where L is the likelihood, and n_{para} is the number of free parameters in the model.) is used. The estimations with a v-copula model have smaller AIC values than the estimations with a Gauss-copula model in all cases, indicating that the v-copula models provide a better fit. Compared with the rank correlations from the Gauss-copula model, the rank correlations of the v-copula model are closer to the empirical rank correlations (Figure 5.11). These results demonstrate the existence of non-multi-Gaussian spatial dependence at the MADE site in high dimensions (where dimensionality here refers to the size of the data subsets). Compared with the bivariate measures in Section 5.3.1, these model fits show that the two datasets exhibit notable differences in higher dimensions (Table 5.3 and 5.4), with the following observations: a) the flowmeter dataset exhibits a more Gaussian type of dependence than the DPIL dataset, which is indicated by the larger value of m_c for the flowmeter data and is also apparent in Figure 5.11; b) the flowmeter dataset has a smaller range in the vertical direction and larger ranges in the horizontal direction than the equivalent isotropic case; c) the Gauss-copula model estimates resulted in smaller ranges than the v-copula fits. The ratios of the horizontal range and the vertical range of v-copula fits (8.36 for DPIL and 14.27 for flowmeter) are closer to the empirical anisotropy ratio (Table 5.2) than the ratios of Gauss-copula fits (10.47 for DPIL and 18.65 for flowmeter); (d) The value of m_c for the flowmeter data in the isotropic case is significantly different from the values estimated in the vertical and horizontal cases; (e) The diversity of the results show the uncertainty of two datasets in different directions.

Table 5.3: Representative parameter estimates for the **v-copula model** based on 200 randomly sampled data subsets of size 10, with the related Akaike information criterion (AIC).

Parameter	DPIL (Ver.)	DPIL (Hor.)	DPIL (Iso.)	Flowmeter (Ver.)	Flowmeter (Hor.)	Flowmeter (Iso.)
V-copula m_c [-]	1.33	1.29	1.50	1.91	1.95	1.41
V-copula k_c [-]	2.75	2.65	2.74	3.09	3.51	2.43
Matern range [m]	3.88	34.57	17.09	3.12	44.53	40.92
Matern κ [-]	0.71	0.34	0.79	0.43	0.20	0.45
AIC [-]	-5562.87	-3166.46	-4901.32	-2969.11	-1598.68	-2805.82

Table 5.4: Representative parameter estimates for the **Gauss-copula model** based on 200 randomly sampled data subsets of size 10, with the related Akaike information criterion (AIC).

Parameter	DPIL (Ver.)	DPIL (Hor.)	DPIL (Iso.)	Flowmeter (Ver.)	Flowmeter (Hor.)	Flowmeter (Iso.)
Matern range [m]	1.51	15.81	8.95	2.31	43.08	20.78
Matern κ [-]	0.79	0.29	0.94	0.42	0.18	0.50
AIC [-]	-5183.33	-2951.80	-4697.32	-2929.49	-1558.70	-2762.78

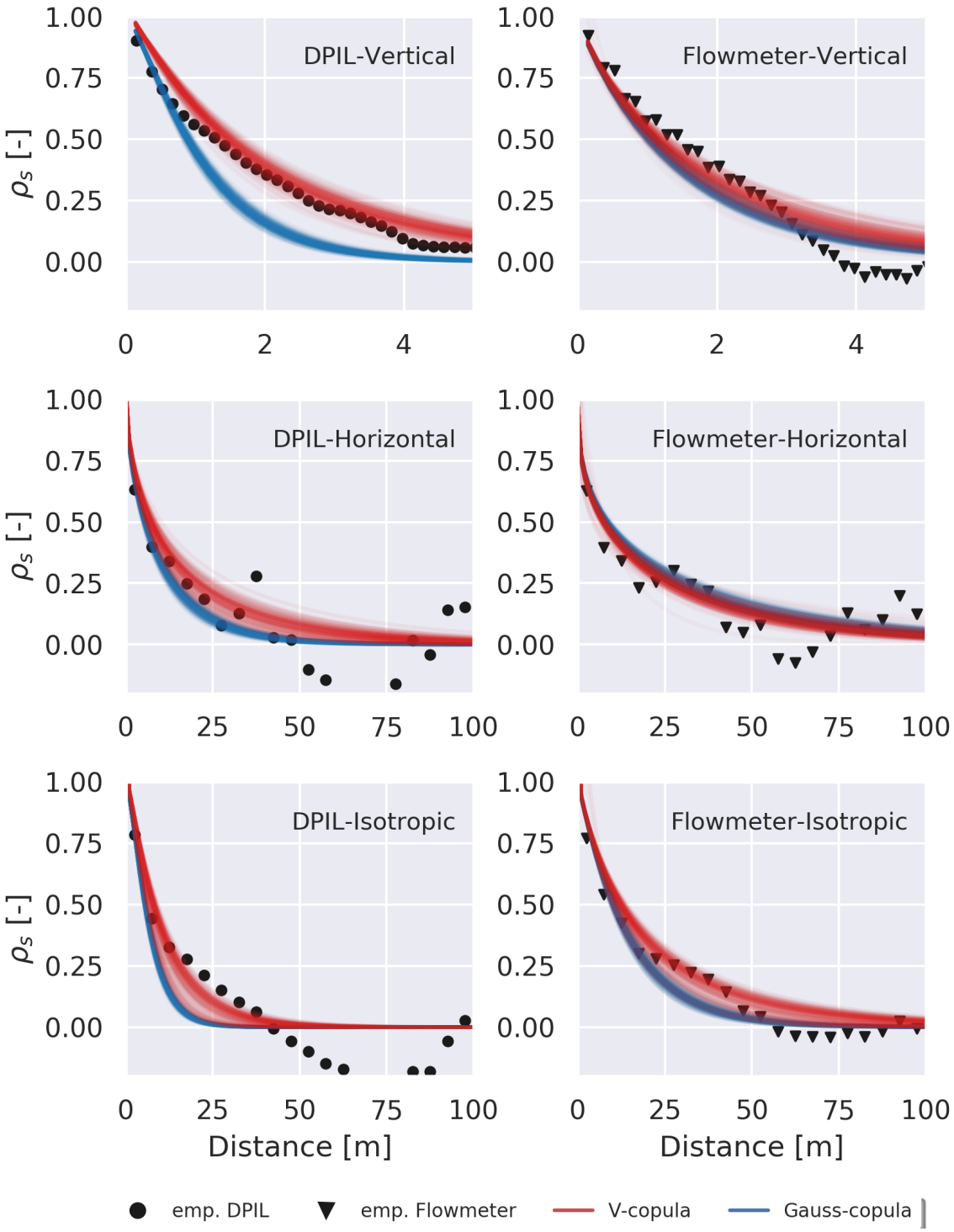

Figure 5.11: Comparison between empirical rank correlation (black dots) and the theoretical bivariate copula rank correlations of model fitting results by Gauss- (blue line) and v-copula (red line) model of DPIL (left) and flowmeter dataset (right).

Possible reasons for the observed differences in parameter estimates include: (1) The behavior of the selected n-points subsets is influenced by varying uncertainty in the measurements of both methods and the different patterns of measurement errors of the two datasets. (2) The distribution of the measurement locations is different between the two datasets. Most notably, the lateral distribution of the flowmeter profiles is more extensive and uniform than that of the DPIL profiles.

5.3.3 Parameter Estimation with Censored Data

Generally, it is expected that any measurement technique performs optimally in a central range of measurement values and suboptimally when the measured values are extremes. Particularly, it has been observed (Bohling *et al.*, 2016) that the DPIL data cannot resolve relatively large K values, as above a certain K value of the formation, the injection-induced pressures are too small to be measured accurately. By contrast, the flowmeter does not resolve hydraulic conductivity below a lower limit because the estimated conductivity is proportional to the gradient of the total flow rate within the borehole (Rehfeldt *et al.*, 1992). However, such relatively large and small K values should not be discarded, because they do carry information, namely, that the unknown true value is somewhere in the interval between a threshold and the largest possible value or the lowest possible value. Such intervals can be included in the estimation by the maximum likelihood method (Section 2.4.3).

Two applications demonstrating the impact of accounting for censored data in the copula analysis are presented, one with a simulated log-normal Gaussian field to prove the concept, and one with real MADE datasets.

In the simulation study, a two-dimensional grid (100×100 cells) was simulated with a known covariance structure ("truth", $\ln(K)$ marginal, mean = 10.0, standard deviation = 6.0, variogram model: $1.0\mathrm{Mat}(5.0)^{0.4}$). A sample of size 2500 was generated and an independent measurement error ε ($\Phi(0.0, 2.0^2)$) was added.

Based on this sample, parameter estimation was performed with the following scenarios: (1) the sample (with ε) as fully certain, (2) the smallest 4%, 8%, 10% and 14% of the samples as being censored and parameter estimation was performed using Equation 2.45. Scenario (2) was established because ε disturbs the rank of smaller samples more than the ranks of larger samples. Treating the smallest 10 percentage of points as censored provided rank correlation estimates that were closest to the virtual truth; this case is compared to scenario (1) (no censoring, (0, 100)) in Figure 5.12. This test shows that the virtual truth and the sampled data (True and True_sample) can be better estimated by including some censorship (i.e., some values below the detection limit) in the copula parameter

estimation when the true value is perturbed by a Gaussian error (True_sample+Err).

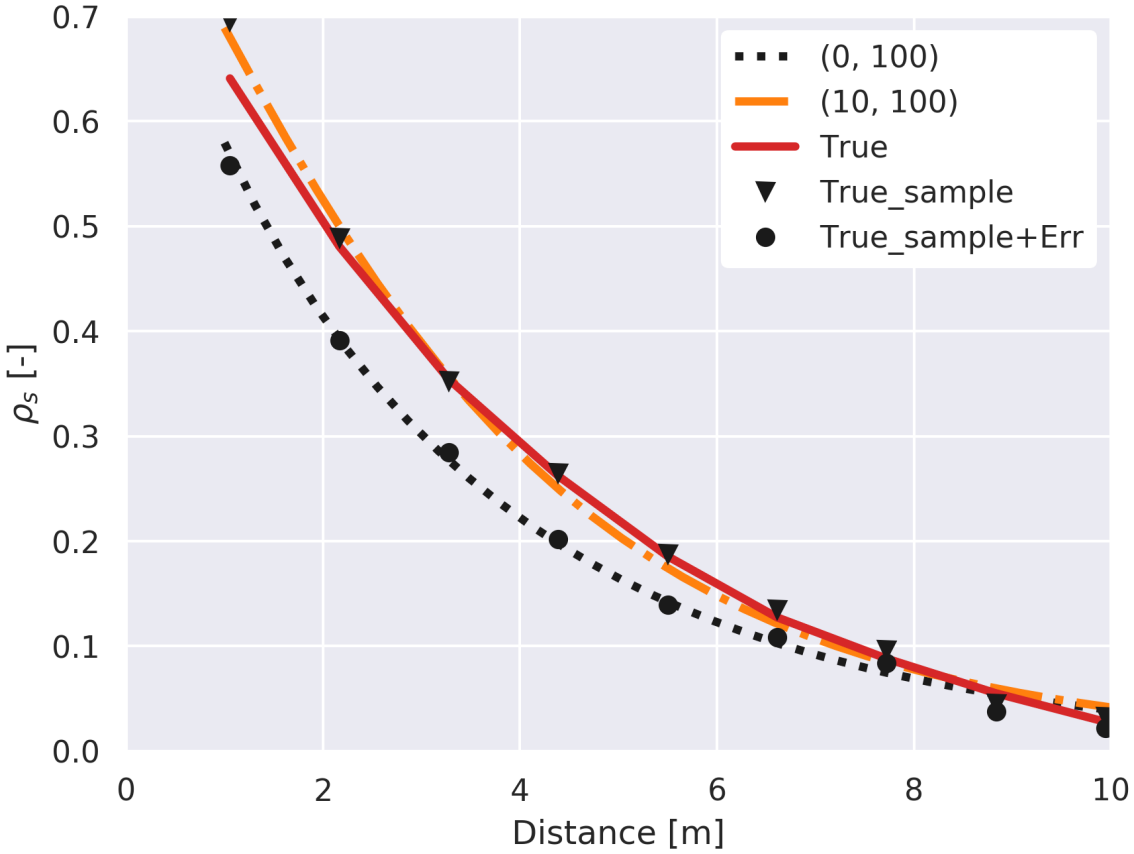

Figure 5.12: Comparison between the Gaussian copula parameter estimation with a 10% left-side detection limit (orange dash-dotted lines with label "(10,100)"), without the detection limit (black dot line with label "(0,100)") and the rank correlations of true domain with 10,000 points (red solid line with label "True"), a sample of the true domain with 2500 points (black inverted triangles with label "True_sample") and the sample with Gaussian noise (black dots with label "True_sample+Err") by using a simulated log-normal Gaussian field.

Unlike in the synthetic test case, the true value of the rank correlation is unknown when using the field data of the MADE site. Therefore, the empirical rank correlation is used as the next best comparable benchmark. According to Figure 5.13, the empirical rank correlation can be better reproduced using the right-side censoring threshold with a censoring threshold of $(0,96)$, which means that the largest four percentage observations are marked as observations with low reliability. For the flowmeter dataset, the empirical rank correlation can be better reproduced by adding a two percentage left-side threshold (Figure 5.14). The estimated rank correlations are calculated by a Monte-Carlo simulation

based on the thresholds from the parameter estimation. Only very small changes were found when the thresholds were small. In summary, the results of parameter estimation can be improved meaningfully when at least some measurements are censored, for the flowmeter data when some small K values were treated as censored and for the DPIL data when some large K values were treated as censored.

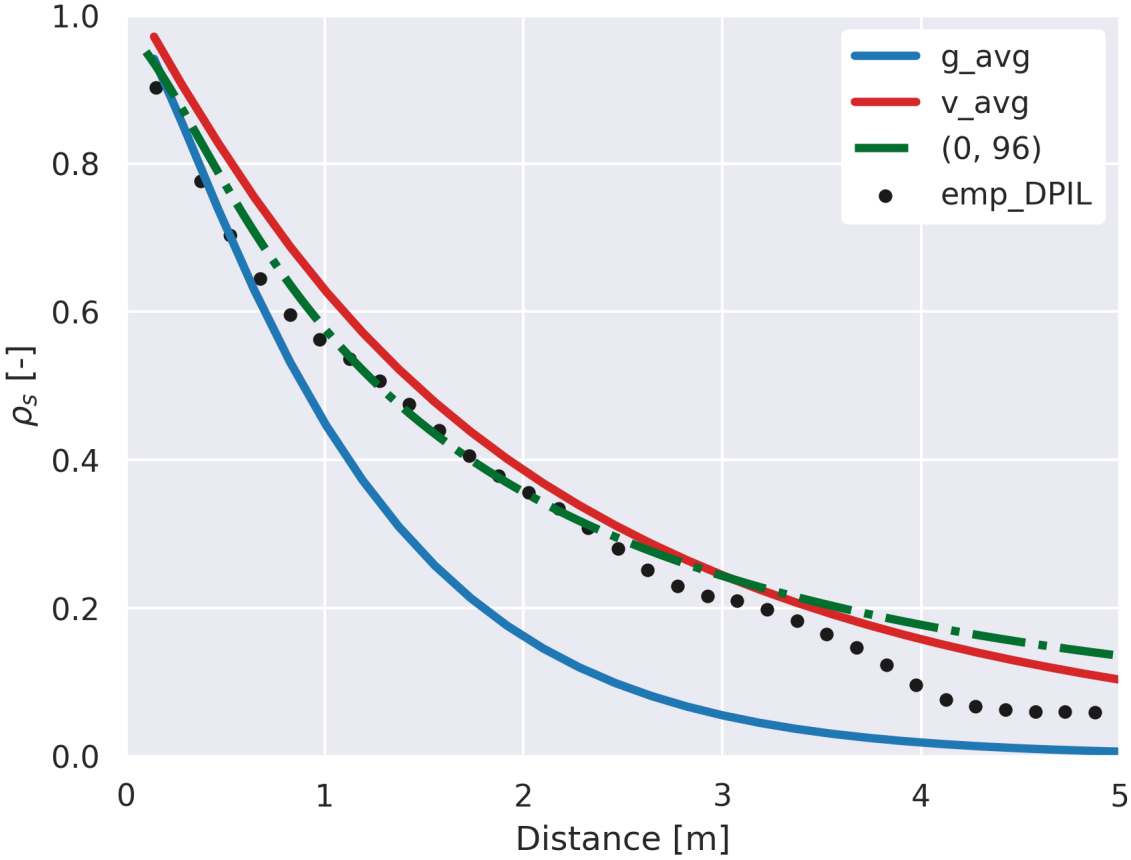

Figure 5.13: Comparison between the Gaussian copula parameter estimation with 4% right side detection limits (dash-dotted green linewith label (0, 96)), without detection limit (solid blue line with label g_avg), v-copula parameter estimation without censored threshold (solid red line with label v_avg) and empirical rank correlation of the DPIL dataset in vertical direction (black dots with label emp_DPIL).

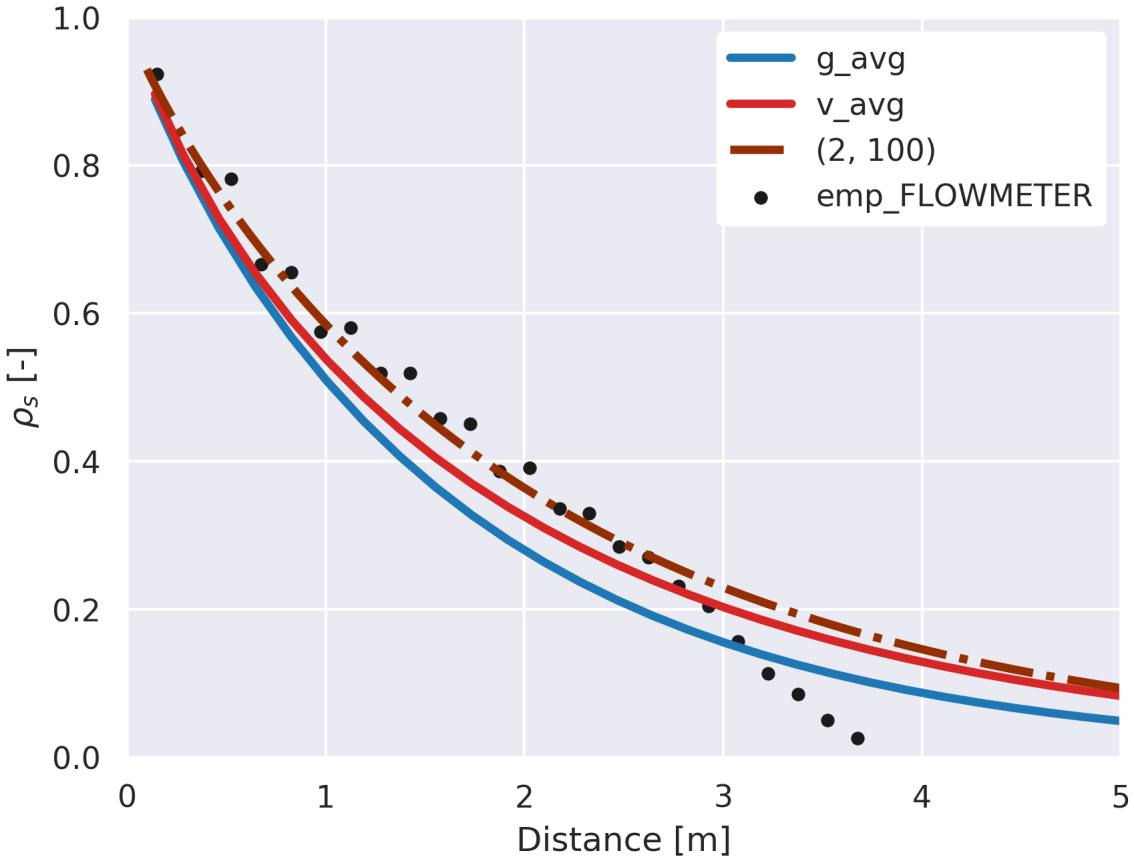

Figure 5.14: Comparison between the Gaussian copula parameter estimation with 2% left side detection limits (dash-dotted brown line with label (2, 100)), without detection limit (solid blue line with label g_avg), v-copula parameter estimation without censored threshold (solid red line with label v_avg) and empirical rank correlation of the flowmeter dataset in vertical direction (black dots with label emp_FLOWMETER).

5.4 Summary and Conclusion of this Chapter

This chapter focuses on the analysis and characterization of the spatial dependence structures of K observations of DPIL and flowmeter dataset at the MADE site using copula-based measures.

The similarity of the two datasets in the marginal distribution (Figure 5.2) and second-order dependence measures are confirmed (Figure 5.7 and (Bohling *et al.*, 2012)). Although in certain separation distances, this similarity is influenced by the lateral coverage of the two datasets (Figures 5.7 and 5.10).

A deviation of both datasets from multi-Gaussian dependence is found, both in empirical measures and mathematical models. The empirical asymmetry (A) systematically and within meaningful confidence intervals deviates from zero indicating that both datasets exhibit non-multi-Gaussian dependence (Figures 5.7 and 5.8). Even more, the strongest dependence can exist in the small K values ($A < 0$) for some lag distances and in large K values ($A > 0$) for other lag distances (Figures 5.8 and 2.11). Such behavior has been found at the MADE site and other relevant hydrogeologic study sites such as Borden, Cape Cod, and North Bay. Together with anisotropy, such a varying degree of dependence in different lag distances will lead to a meaningfully improved description of solute transport as it describes loosely speaking barriers to flow and connected high-K channels.

The rank correlations of the two datasets between horizontal lags of 2.5 [m] and 17.5 [m] are more similar to each other than are the variogram values for these lags (Figure 5.7), possibly due to the reduced influence of extreme values on the rank correlation. The datasets also have similar bivariate copula densities (Figure 5.9). For most lags, particularly for the short separation distances that are important when analyzing solute transport, the DPIL and flowmeter bivariate copula densities are not deemed to be significantly different, based on a χ^2 test with a significance level of $\alpha = 0.05$ (Figure 5.10).

The parameters of two different theoretical copula models were estimated for both datasets. The multi-Gaussian model and a non-multi-Gaussian model (v-copula) were used. The v-copula model provided a better fit to both datasets (Figure 5.11) than the Gauss-copula, indicating meaningful non-multi-Gaussian spatial dependence.

When the uncertainty of censored measurements is incorporated into the multi-Gaussian model, the resulting parameter estimates are closer to reality than when they are not included and are also similar to the estimates obtained with the v-copula model without censored measurements (Figures 5.13 and 5.14).

This indicates that some of the censorship properties might be mimicked by the shape of the v-copula. The ability to obtain improved Gauss-copula model fits after accounting for censoring does not conflict with the conclusion that the spatial dependence is non-multi-Gaussian, which is indicated not only by the rank correlation but also the asymmetry.

This research identified two criteria that a realistic geostatistical model for spatially distributed K fields should fulfill: (1) it should be able to express a varying uncertainty of measurements, and (2) it should more adequately resemble the processes that lead to the empirical and the theoretical measures from the observations. One important feature of an improved resemblance was identified to be a model that leads to both positive and negative values of the copula asymmetry for varying lag distances.

Chapter 6

Hydraulic Conductivity Simulation and Evaluation of Macrodispersion

In this chapter, the phase-annealing and the particle-tracking random-walk method are applied to the conditions at the MADE site in order to model solute transport using K fields based on different geostatistical models. The setup of the simulations is presented in Section 6.1 and the simulation results are analyzed in Section 6.2. A summary of this chapter can be found in Section 6.3.

6.1 Application to the MADE Site

A shifted MADE-2 coordinate system (see Boggs *et al.*, 1992, and Figure 6.1) is used in the simulations, in which x_2 is the coordinate along the direction of predominant flow (longitudinal direction) and x_3 is the vertical direction. To harmonize the vertical resolution of the DPIL (0.015 [m]) and flowmeter (0.15 [m]) datasets, ten DPIL observations in the vertical direction are averaged before using it in the conditional simulations. In the following, the details of the asymmetry accounted for in the simulation (Section 6.1.1), the choice of primary and secondary datasets in the conditioning procedure (Section 6.1.2), as well as the configurations of the phase-annealing simulations and the particle-tracking random-walk simulations (Section 6.1.3) are presented.

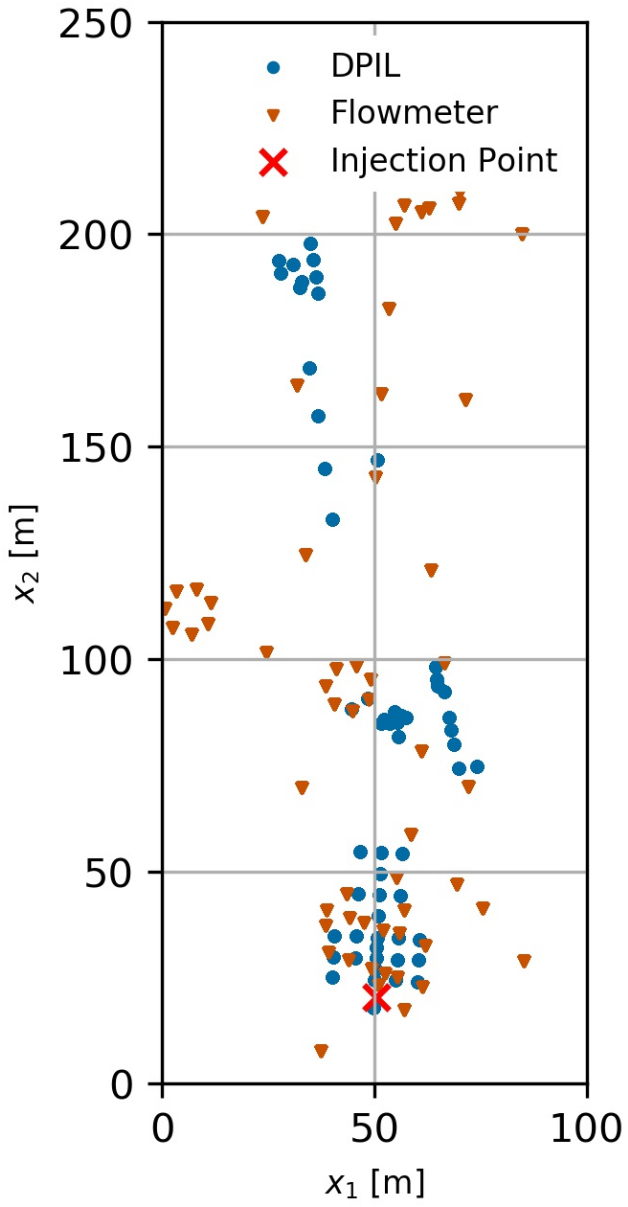

Figure 6.1: Plan view of the domain. Orange triangles: locations of flowmeter tests, blue circles: locations of direct-push injection-logging (DPIL) observations, red cross: injection point of the tracer test. Mean flow is oriented in the x_2 direction.

6.1.1 Assessment of Global and Depth-Dependent Asymmetry

The copula-based asymmetry A in Equations 2.27 and 2.28 provides third-order spatial information of the spatial non-Gaussianity of the MADE dataset going beyond the second-order variogram-based geostatistics. Guthke and Bárdossy (2017) determined the asymmetry for the whole domain, either as an isotropic or anisotropic property. This is denoted as "global asymmetry A".

At the MADE site, the horizontal asymmetry varies significantly with depth (Figure 6.2A). Data pairs on the top (close to the ground surface) show more positive asymmetries and data pairs at the bottom exhibit more negative asymmetry, which is somewhat averaged out when assessing the "global" asymmetry A.

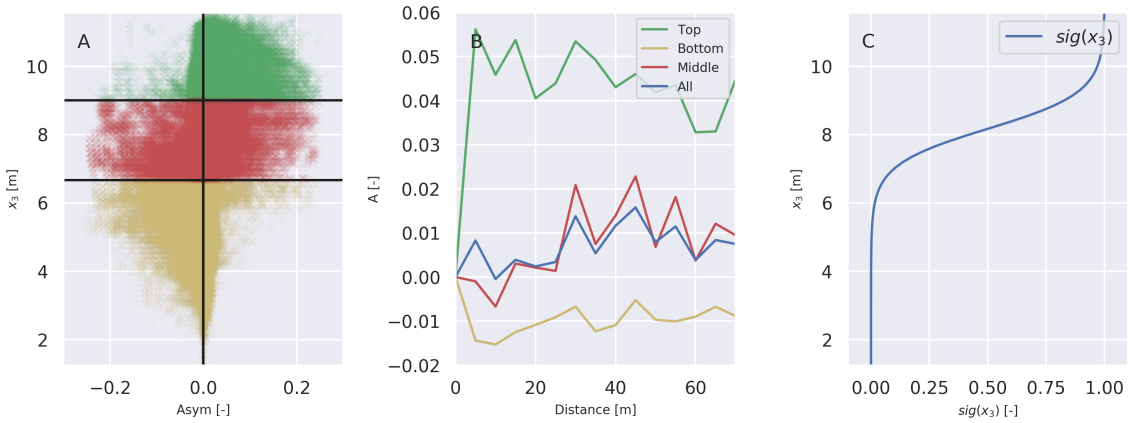

Figure 6.2: Analysis of the copula asymmetry (A) of the flowmeter data. A) Contributions of pairwise horizontal asymmetry as a function of depth. Colors indicate an association with three distinct zones; pink: top, green: middle, blue: bottom. The boundaries between the layers are located at $x_3 \sim 6.67\,[m]$ and $x_3 \sim 9.0\,[m]$. B) Horizontal empirical asymmetry of the flowmeter dataset in the top, middle, and bottom layers, and averaged over the entire depth. C) Vertical profile of the weighting function $\mathrm{sig}(x_3)$.

This observation is accounted for in two alternative ways within our simulations: In one approach, three distinct layers (with boundaries $x_3 = 6.67\,[m]$ and $x_3 = 9.0\,[m]$) with meaningfully different asymmetry A (*Hard A*) are defined by maximizing the difference of the Asymmetries between the top and the bottom layer, whereas in the other a continuous function $A(x_3)$ of the asymmetry A (*Smooth A*) as a function of the vertical coordinate x_3 is developed.

Figure 6.2B shows the empirical global asymmetry and asymmetries for the three distinct layers. The global horizontal asymmetry is mild, whereas the empirical horizontal

asymmetry of the top layer is positive for all separation distances and that of the thicker bottom layer is slightly negative. The middle layer has an asymmetry function similar to the global estimate. If the layer-dependent asymmetries are accounted for in the conditional simulations, the resulting K fields reproduce the structure of the flowmeter dataset better. In this case, the phase-annealing iteration searches in a more compact space than when using the global asymmetry function. The latter also reduces the required number of iterations in the Monte-Carlo simulation.

In the alternative approach of accounting for the depth dependence of the asymmetry, a smooth function $A(\hat{x_3})$ is defined that gradually varies between the top (A_{top}) and bottom (A_{bottom}) asymmetry using a sigmoid weighting function $\mathrm{sig}(x_3)$ plotted in Figure 6.2C:

$$A(\hat{x_3}) = \mathrm{sig}(x_3)A_{top} + (1 - \mathrm{sig}(x_3))A_{bottom}, \tag{6.1}$$

with

$$\mathrm{sig}(x_3) = \frac{1}{1 + \exp(-(a \cdot x_3 + b))}, \tag{6.2}$$

in which $a = 1.89[1/m]$ and $b = -8.08$ are two fitted parameters.

The asymmetry function $A(\mathbf{h})$ in Equations 2.27 and 2.28 is point-symmetric with respect to the separation vector $\mathbf{h}$, i.e. $A(\mathbf{h}) = A(-\mathbf{h})$. This implies that the effects of the distinct layers disappear in the vertical asymmetry when points belonging to two different layers are considered. Only the horizontal asymmetry is considered in the objective function obj_3 to push large values on top, A possible equivalent approach would be to use a directional vertical asymmetry $A(-\mathbf{h}) \neq A(\mathbf{h})$ (Bárdossy and Hörning, 2017). Another reason for using the horizontal depth-dependent asymmetry A lies in the unbalanced distribution of the observation points in the main flow direction (Figure 6.1). The vertical resolution of the flowmeter and DPIL logs are so high that the K profiles are strongly conditioned in the vertical direction in the direct vicinity of the measurement profiles, no matter which statistical metrics are included in the simulation. Between the measured profiles, by contrast, the uncertainty of the simulated fields can significantly be reduced by accounting for an accurate description of statistical dependence in the horizontal direction.

6.1.2 Choice of Primary and Secondary Information

While both the flowmeter and DPIL dataset yield estimates of local hydraulic conductivity values at the MADE site, they are based on different measurement principles that are limited in their accuracy in different ranges of K. This leads to different marginal distributions but similar spatial dependence structures (Bohling *et al.*, 2012; Xiao *et al.*, 2019,

and Chapter 2). Rather than trying to transform the measurements of one investigation technique, e.g., the Q-Q transformation, to match the marginal distribution of the other, one dataset is used as the "primary" and the other as the "secondary" dataset to include both datasets in one geostatistical model. The flowmeter dataset is chosen as a primary variable because the flowmeter observations cover a larger area and are more spread out in space than the DPIL observations (Figure 6.1). The DPIL dataset (averaged to the same vertical resolution as the flowmeter data) is used as a secondary variable. From the primary dataset, the geometric mean (4.29×10^{-5} [m/s]) and variogram of the K-field are taken (exponential with correlation lengths of $12.3m$ in the horizontal and $1.5m$ in the vertical direction, variance of $\log(K) = 4.41$), and the realizations are conditioned on the point values of this dataset.

While the two datasets differ in their marginal distributions (Figure 5.2), they show high similarity in their rank correlation functions (second row in Figure 5.7). Therefore, the K fields are conditioned not on the actual values of the DPIL dataset but the rank (order of point values) within the measured values. This semi-quantitative information (large versus small K value within the set) is independent of the measurement tools applied. In this thesis, a zonation of the mean or variogram in the simulation is not considered, even though geostatistical tools exist to identify the boundaries of zones with different properties (Haslauer *et al.*, 2017a).

6.1.3 Tested Types of Simulation

Several geostatistical models that differ in the information included in the simulations are tested. All models use the same marginal distribution and variogram. Table 6.1 gives an overview of all models, in which each row represents a different geostatistical model. The set of the model includes:

g_0: The base case included for reference is an unconditional multi-Gaussian simulation.

g_1: The simplest conditional model is a multi-Gaussian simulation honoring the point values of the flowmeter measurements.

g_2: The next model is a multi-Gaussian model honoring the point-values of the flowmeter data and the ranks of the DPIL data.

v_{11}: This non-multi-Gaussian model is conditioned on the point-values of the flowmeter measurements, but not on the DPIL data, and assumes the asymmetry functions determined by the three distinct layers discussed above.

v_{12}: This non-multi-Gaussian model is identical to v_{11} but is additionally conditioned on the ranks of the DPIL measurements.

v_{21}: This non-multi-Gaussian model assumes the smooth transition of asymmetry from the top to the bottom, is conditioned on the point-values of the flowmeter-derived K values, but neglects the DPIL measurements.

v_{22}: may be seen as the most complex model, accounting for the smooth trend in the asymmetry function, honoring the point values of the flowmeter measurements and the ranks of the DPIL measurements.

Table 6.1: Information included in the different simulation models. Hard A: depth-dependent asymmetry function in three distinct layers; smooth A: asymmetry function that smoothly varies with depth.

Geostat. Model	Variogram	Flowmeter Value	DPIL Rank	Hard A	Smooth A
g_0 —	✓				
g_1 —	✓	✓			
g_2 —	✓	✓	✓		
v_{11} —	✓	✓		✓	
v_{12} —	✓	✓	✓	✓	
v_{21} —	✓	✓			✓
v_{22} —	✓	✓	✓		✓

For each model, $N = 100$ realizations of the three-dimensional K fields are simulated. The properties of the simulated K fields are analyzed and compared between different simulations in the following sections.

6.1.4 Set-Up of Flow and Transport Simulations

Table 6.2 lists the configurations of the K simulations and the flow-and-transport calculations. The domain has a dimension of $\mathbf{L} = 100\,[m] \times 250\,[m] \times 15\,[m]$ and is discretized by $\Delta\mathbf{x} = 1.0\,[m] \times 1.0\,[m] \times 0.15\,[m]$. Each grid cell has a different conductivity value. In the flow simulations, a mean hydraulic gradient of 3×10^{-3} is applied by constant-head boundary conditions on the boundary faces of the domain perpendicular to the x_2 direction. A uniform porosity of 0.35 is assumed (Adams and Gelhar, 1992). In each realization, 10,000 particles are released at the point $\mathbf{x_{inj}} = (50\,[m]; 20\,[m]; 7.5\,[m])$. A uniform transverse dispersion coefficient of $D_t = 1 \times 10^{-8}\ m^2/s$ is applied avoiding issues related to the gradient of the dispersion coefficient in the random-walk simulations. The chosen transverse dispersion coefficient is about ten times larger than the typical values of the pore diffusion coefficient D_p.

Table 6.2: Configurations of K fields simulation and particle tracking simulation.

Simulation of K-Field

Vertical correlation length	l_v	$[m]$	1.5 (Rehfeldt *et al.*, 1992)
Horizontal correlation length	l_h	$[m]$	12.3 (Rehfeldt *et al.*, 1992)
Domain size	$\mathbf{L}$	$[m]$	$(100; 250; 15)$
Grid Spacing	$\Delta\mathbf{x}$	$[m]$	$(1.0; 1.0; 0.15)$

Particle-Tracking Random-Walk Simulation

Mean hydraulic gradient	i	[-]	3×10^{-3} (Boggs *et al.*, 1992)
Porosity	n	[-]	0.35 (Adams and Gelhar, 1992)
Transverse Dispersion Coefficient	D_t	$[m^2/s]$	1×10^{-8}
Number of particle	N_p	[-]	1×10^4
Injection location	$\mathbf{x_{inj}}$	$[m]$	$(50; 20; 7.5)$

6.2 Results and Discussions

The statistical properties of the simulated K fields of different geostatistical models in Table 6.1 are analyzed in Section 6.2.1 and the statistical analysis of the particle tracking results is presented in Section 6.2.2.

6.2.1 Statistical Analysis of Simulated K Fields

To give a visual impression of the K fields simulated by the different models, Figure 6.3 presents horizontal cross-sections ("plan view") through a single example field of each method at three different depths (top: $x_3 = 13.5\,[m]$, middle: $x_3 = 8.4\,[m]$, bottom: $x_3 = 3.0\,[m]$). Rather than showing values of the conductivity, Figure 6.3 presents the corresponding cumulative probability values $F_z(K)$.

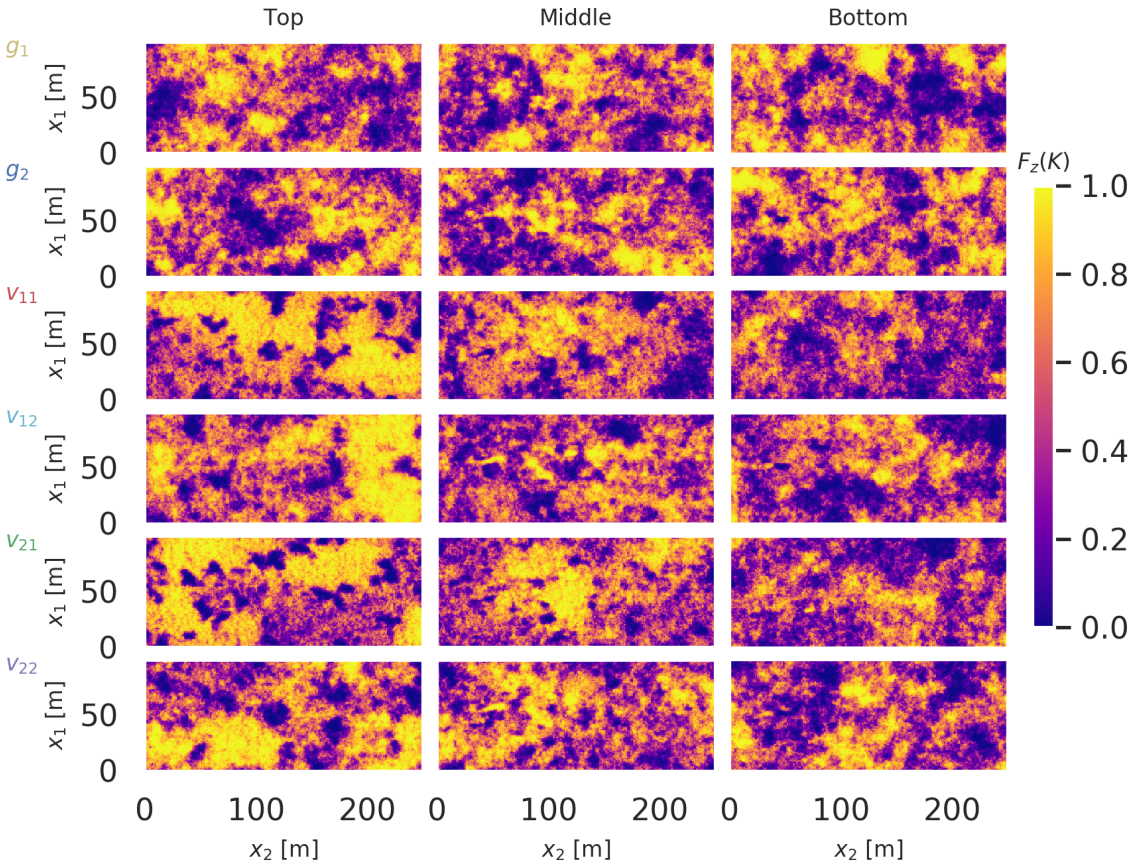

Figure 6.3: Plan views of example realizations for each simulation model of the simulated K-field in empirical distribution function space ($F_z(K)$); Top: horizontal cross-section at $x_3 = 13.5\,[m]$, middle: cross-section at $x_3 = 8.4\,[m]$, bottom: cross-section at $x_3 = 3.0\,[m]$.

The fields of the two conditional multi-Gaussian models, g_1 and g_2 are structurally similar and don't show any systematic differences in the $F_z(K)$-patterns with depth. When the depth-dependent asymmetry is accounted for in the conditional simulations (v_{11}-v_{22}), a systematic local structure is introduced. There are larger isolated patches ("blobs") of high-K values in the top cross-section than in the bottom cross-section. This pattern is

clear in the plot of the top column of the simulation v_{11}-v_{22} (Figure 6.3).

In the following, the cumulative distributions of K at individual points $\mathbf{x}$ are compared to the marginal distribution $F_z(K)$ that went into the conditional simulations as a global property. All K values are expressed in terms of their $F_z(K)$ values, making the analysis independent of the marginal distribution. The probability that the local value exceeds a given value z at location $\mathbf{x}$ is denoted $P_{ens}(z(\mathbf{x}))$. The corresponding empirical distribution functions at individual points can be characterized by their ensemble mean and variance, yielding point measures of the distribution function simulated by different methods.

With a sufficiently larger ensemble size, an ensemble of unconditional multi-Gaussian realizations yield an empirical distribution that is identical to the marginal distribution $P_{ens}(z) = F_z$ with the ensemble point mean $\langle z \rangle = 0.5$ and ensemble point variance $\sigma^2(z) = \frac{1}{12}$ at all locations $\mathbf{x}$. Conditioning leads to deviations between $P_{ens}(z)$ and F_z, which may be expressed as:

$$P_{ens} = F_z + P_0'(z(\mathbf{x_i})) + P_1'(z(\mathbf{x_i}), \gamma) + P_2'(z(\mathbf{x_i}), \gamma) + P_3'(z(\mathbf{x_i}), \gamma, A), \tag{6.3}$$

in which P_0' is a perturbation already occurring in the unconditional simulation (due to an insufficiently large ensemble size), P_1' is a perturbation introduced by conditioning to point values of the primary variable, P_2' is a perturbation introduced by conditioning to the orders of the point values of the secondary variable, and P_3' is a perturbation caused by honoring the asymmetry of the copula.

Conditioning to point values and orders of the point values fixes the simulated values at the conditioning points within the measurement error of the observations. In the direct vicinity of these observation points, the variability is reduced, which is expressed in Equation 6.3 by the dependence of P_1' and P_2' on the variogram function γ, and of course on the distance to the observation points.

The perturbation caused by the asymmetry term P_3' depends on the asymmetry A. According to Equations 2.27 and 2.28, a positive asymmetry leads to a larger P_{ens} on large values and a negative asymmetry leads to a larger P_{ens} on small values. Accounting for the asymmetry A in the conditional simulations may lead to a reduction or increase of the variance in comparison to the original dataset, unlike in multi-Gaussian conditional simulations, the effect on the variance depends not only on the distance to the observation points but also on the actual values observed.

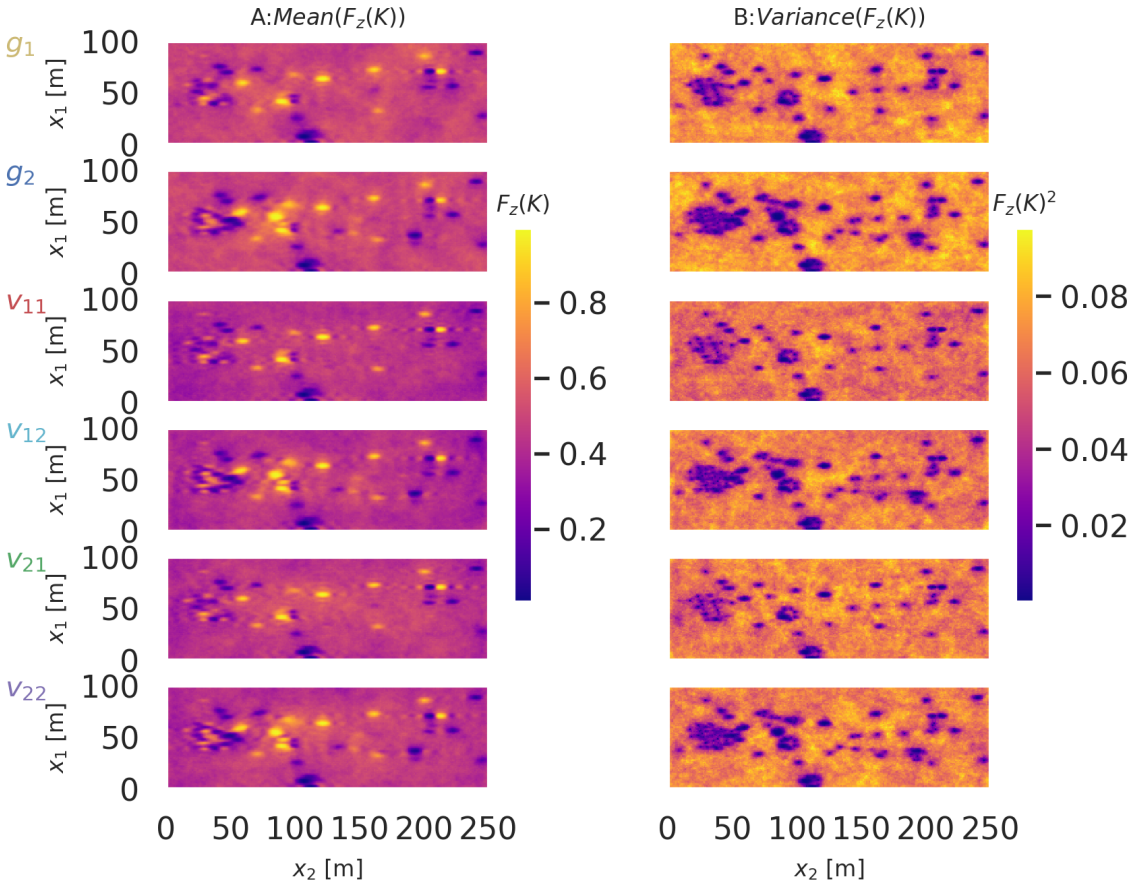

Figure 6.4: Horizontal cross-sections of the ensemble point mean and variance of $F_z(K)$ at $x_3 = 7.5\,[m]$ for the different simulation models. A: ensemble point mean of $F_z(K)$; B: ensemble point variance of $F_z(K)$.

Figures 6.4A and 6.4B present the horizontal cross-sections of the ensemble point mean and variance, respectively, of $F_z(K)$ at $x_3 = 7.5\,[m]$ for the different conditional simulations (g_1-v_{22}) to show the influence of the different information included. Like in Kriging, the variances are significantly reduced at observation points and their direct vicinity. This holds for both the point-value observations of the primary flowmeter data and the point-order observations of the DPIL measurements (see additional spots of reduced variance in models g_2, v_{12}, and v_{22} in comparison to models g_1, v_{11}, and v_{21} in Figure 6.4B). This implies that the order information is similarly effective as direct measurement values in reducing the conditional uncertainty of the simulated K fields.

The same pattern can be found in the plots of mean values in Figure 6.4A. Bright and dark points mark the locations of observations with high and low values, respectively. These points generate local features of the spatial K-field. Note that Figure 6.4A shows a

low-K zone at $30\,[m] \lesssim x_1 \lesssim 70\,[m]$ and $20\,[m] \lesssim x_2 \lesssim 60\,[m]$ which is downstream of the injection point $\mathbf{x}_{inj}$ in the MADE-2 experiment. This is a possible reason for the observed small tracer velocity in the source zone (a well-known property of the MADE site, see Zheng *et al.* (2011)).

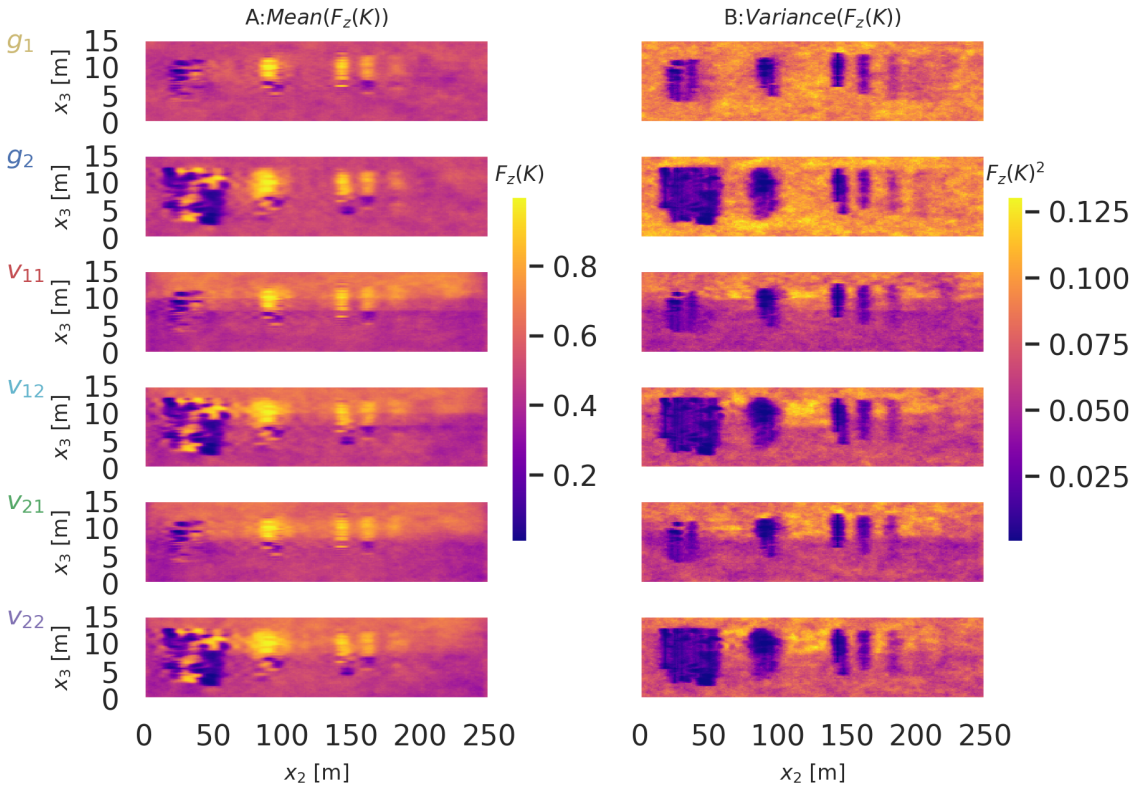

Figure 6.5: Vertical cross-sections of the ensemble point mean and variance of $F_z(K)$ in the direction of mean flow at $x_1 = 7.5\,[m]$ for the different simulation models. A: ensemble point mean of $F_z(K)$; B: ensemble point variance of $F_z(K)$.

Figures 6.5A and 6.5B show vertical cross-sections of the ensemble point mean and variance, respectively, at $x_1 = 50\,[m]$. These vertical cross-sections show the above-mentioned low conductivity zone at $20\,[m] \lesssim x_2 \lesssim 60\,[m]$ much clearer than the horizontal cross-sections of Figure 6.4. An additional low-K zone can be found at $x_2 \sim 60\,[m]$ when the DPIL information is included in the conditional simulations (models g_2, v_{12}, and v_{22}). The additional features detected by the DPIL measurements may influence the results of flow-and-transport simulations.

Figure 6.5 also clearly shows that accounting for the asymmetry in the conditional simulations (models v_{11}, v_{12}, v_{21}, v_{22}) causes different patterns of the mean and variance

in the top and bottom parts of the domain. The strong deviation of the asymmetry from zero leads to a large change in P_{ens} in Equation 6.3. The models accounting for depth-dependent asymmetry show a larger mean and variance at the top in comparison to the bottom. Note that the top layer exhibited a particularly large positive asymmetry causing an extended influence of large values. The multi-Gaussian conditional simulations also show several spots of high K values in the top; their spatial influence on the mean, however, is limited, and the reduction of the variance is independent of the vertical position and measured value.

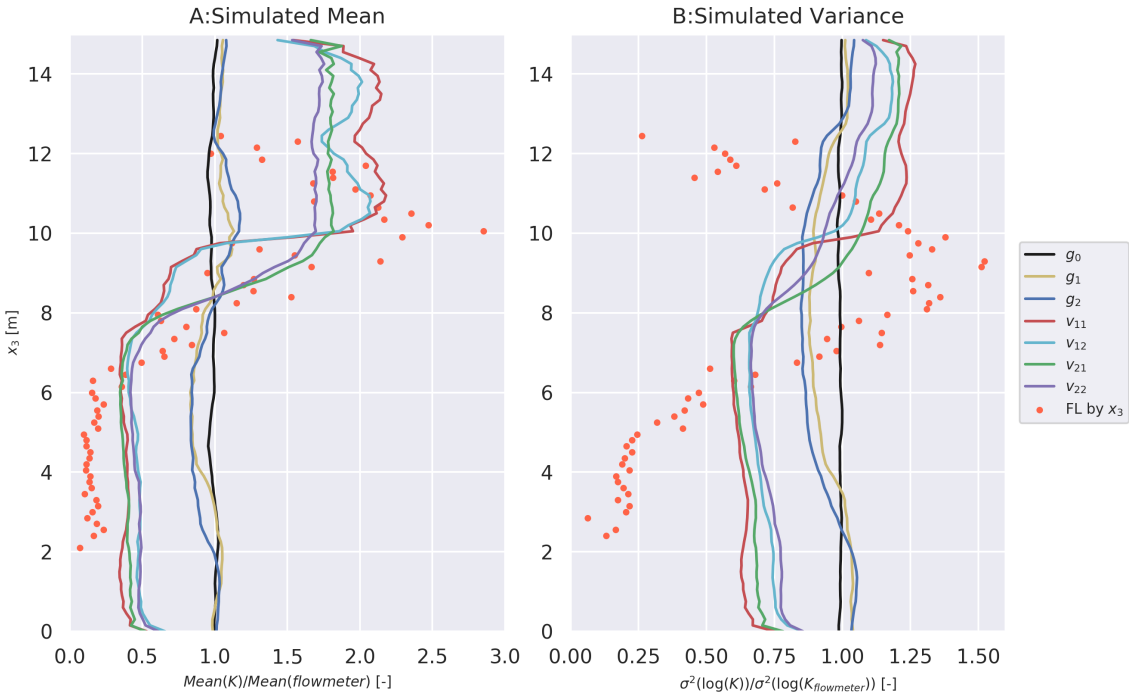

Figure 6.6: Vertical profiles of normalized log-hydraulic conductivity. A: mean, B: variance of $\ln(K)$. Markers: flowmeter data; lines: horizontal averages of the various simulation models.

Figure 6.6 shows vertical profiles of the mean and variance of the flowmeter-derived normalized conductivity values in comparison to the corresponding profiles according to the different geostatistical simulations. The unconditional Gaussian model (g_0) has a vertically uniform mean and variance, both of which are close to the corresponding global metrics of the flowmeter data. Conditioning on the point values and orders within the multi-Gaussian framework (g_1 and g_2), leads to a slight shift of the mean values towards the depth-dependent mean values of the flowmeter data, but the effect is limited to points in the direct horizontal vicinity of the observation profiles so that the depth profile averaged over the horizontal directions shown in Figure 6.6A hardly differs from

the uniform prior. Within the multi-Gaussian framework, the estimation variances are shifted towards smaller values because conditioning points reduce the uncertainty close to those points, creating hot spots of small estimation variance contrasting to a background of high variances close to the original global variance in Figure 6.4B. The vertical profile of the estimation variance thus directly depends on the number of observation profiles.

Accounting for the depth-dependent asymmetry in the non-multi-Gaussian conditional simulations (model v_{11}-v_{22}) yields vertical profiles of the mean and variance close to the corresponding depth-dependent profiles of the flowmeter data. Especially the models with a smooth depth dependence of the asymmetry A show a gradually varying profile of the mean conductivity, which resembles the depth profile of the horizontally averaged flowmeter data. The asymmetry also causes a distinct vertical trend in the variance, which does not resemble the variance of the flowmeter data too much. The increase in the variance in the top zone caused by the asymmetry is smaller when the DPIL measurements are included in the conditioning, which is an effect of variance reduction by conditioning to observations.

All conditioning methods introduce non-stationarity, as they result in spatial patterns of the mean and variance. Moreover, all methods simulate local features if these were hit by the observation profiles. The non-multi-Gaussian methods, however, lead to fields with horizontally more persistent patterns. It is to be expected that this has a more profound effect on flow and transport than the locally constrained conditioning of the multi-Gaussian approach, which is discussed in the next section.

6.2.2 Analysis of Particle-Tracking Results

How the differences in the simulation models of hydraulic conductivity affect the macroscopic characteristics of solute transport is analyzed in this section. The analysis is mainly based on particle travel times observed at the control planes.

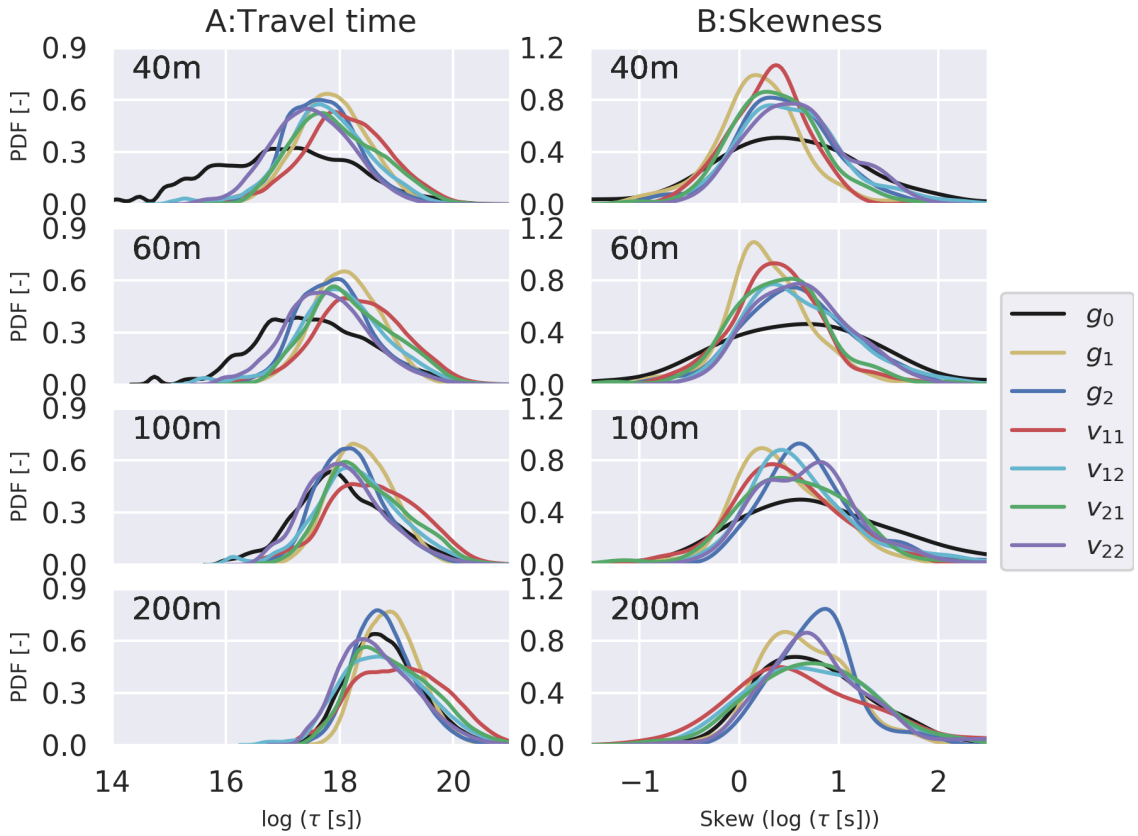

Figure 6.7: Ensemble log-travel time distributions observed in different observation planes perpendicular to the mean direction of flow for the different geostatistical simulation models employed. A: PDF of the ensemble log-travel time; B: PDF of the skewness of log-travel times distributions of individual realizations.

Figure 6.7 shows the logarithmic travel time distributions at four different travel distances for all simulation models employed. While Figure 6.7A shows the distributions of all particles in all realizations belonging to a particular geostatistical model, Figure 6.7B shows the distribution of the skewness of log-travel time within the corresponding ensembles. If macroscopic transport was Fickian, the travel-time distribution would be the inverse Gaussian distribution, which is almost indistinguishable from the log-normal distribution at sufficient travel distance. This implies that the skewness of the log-travel time would be zero. Strong deviations from zero are indicative of anomalous transport.

As listed in Table 6.2, all particles are introduced at $\mathbf{x}_{inj} = (50; 20; 7.5)\,[m]$. Figure 6.5 has shown that all conditional simulations infer a low-K zone shortly downstream of the injection point. This leads to a significant shift towards large travel times from unconditional to the conditional simulations at $x_2 \approx 40\,[m]$ (Figure 6.7A). At larger travel

distances, the effects of conditioning on the overall log-travel time distribution vanishes as the particles now pass through a part of the domain with fewer observations, in which their behaviors are less influenced by conditioning and the global geostatistical properties of the simulated field dominate.

The simulations that account for the copula-based depth-dependent asymmetry A of the K-field ($v11$ - $v22$) show a slightly wider distribution of log-travel time than the multi-Gaussian simulations. However, the only clearly distinguishable model is v_{11} that assumes layers of distinct asymmetry and neglects the DPIL data.

All simulations cause predominantly positive skewness of the log-travel time distribution, which is equivalent to enhanced tailings. This results from the variance of log-hydraulic conductivity.

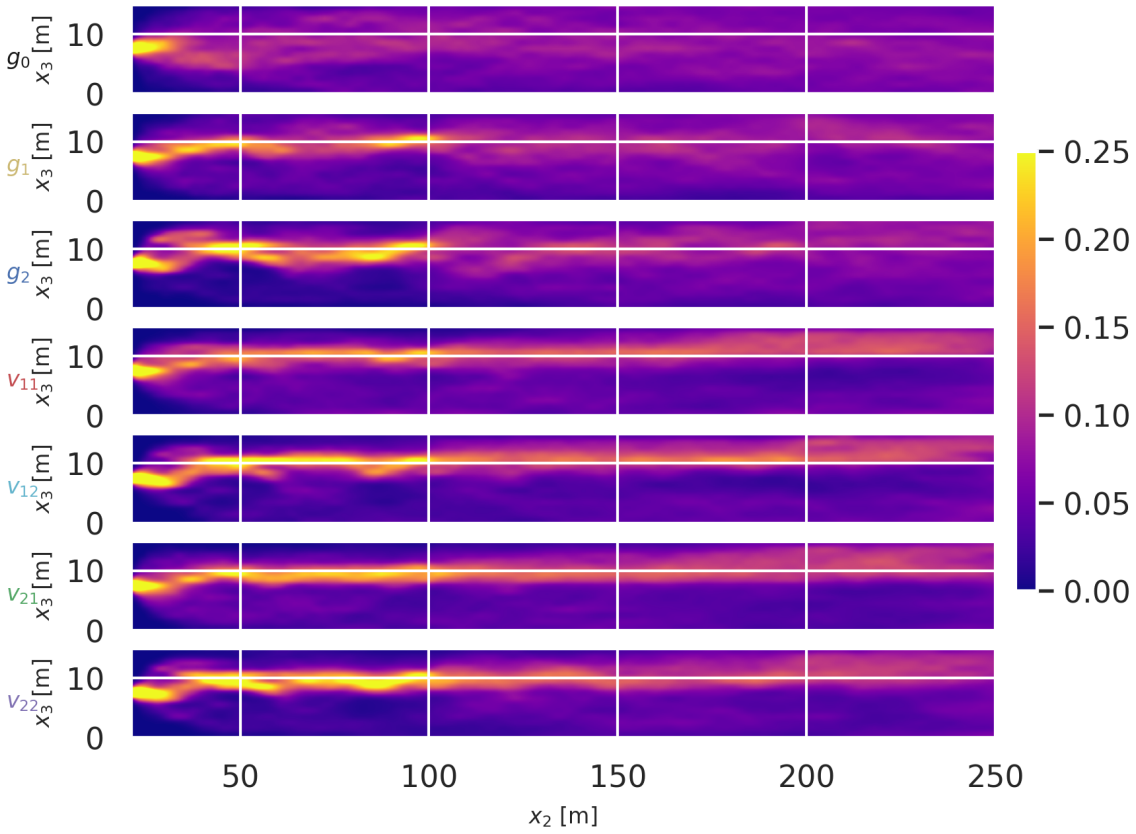

Figure 6.8: Density of the ensemble particle locations in the vertical direction in each observation plane.

Figure 6.8 shows the density of the particle locations in the vertical direction in each

observation plane, exemplifying the introduction of non-stationarity upon conditioning. All particles are released at the same location on the left side. In the unconditional case (g_0), the particles spread symmetrically in the vertical direction, whereas there is a clear upward drift shortly downstream of the injection point in all conditional simulations. Already conditioning on the flowmeter data (g_1, v_{11}, v_{21}) causes the introduction of a low-K zone downstream of the injection point (see Figure 6.5A) that the particles largely try to bypass. Additional conditioning on the DPIL data (g_2, v_{12}, v_{22}) enhances the effect.

In the conditional multi-Gaussian case (g_1 and g_2), the effect of channelizing transport in a preferential flow path in the upper half of the domain vanishes at $x_2 \sim 110\,[m]$ because the density of observation points in this zone is small so that the conductivity and velocity statistics relax to the unconditional case. This is different in the simulations that account for depth-dependent asymmetry (v_{11}-v_{22}). Already the mean conductivity ranks shown in Figure 6.5A and the vertical profiles of mean log-conductivity in Figure 6.6 indicate that these simulations lead to a stronger persistence of high-conductivity values in the top half which affects that most particle trajectories remain in there until the end of the domain.

The existence of a high-conductive layer on top of a less conductive layer controls the overall transport behavior at the MADE site. The particles get stuck in the low-K zone, and as soon as they get out and reach the high-K zone, they stay there and move quickly downstream.

Figure 6.9 shows length profiles of the travel-time derived longitudinal macroscopic velocity (U_ℓ, left column) and dispersivities (D_ℓ/U_ℓ,right column). Solid lines are the effective velocity and dispersivities, in which the transport coefficients are computed for each realization first and averaged over the ensemble afterwards, whereas dashed lines are the ensemble velocity and dispersivities, in which the travel times of all realizations are merged first and the transport coefficients are computed afterwards. The ensemble coefficients thus do not describe how individual plumes behave as they include the variability between the different realizations, which is a metric of uncertainty (Attinger *et al.*, 1999; Dentz *et al.*, 2000). Cirpka (2002) could show that the effective dispersion coefficient for a point injection is a measure of solute mixing, relevant for mixing-controlled reactive transport. In general, the effective dispersion is expected to lag behind the ensemble dispersion (Dentz *et al.*, 2000). This is also what has been observed in our simulations. In contrast to the analytical derivation of Dentz *et al.* (2000), however, the evolution of the dispersion coefficients is affected by the non-stationarity of the conditional conductivity fields.

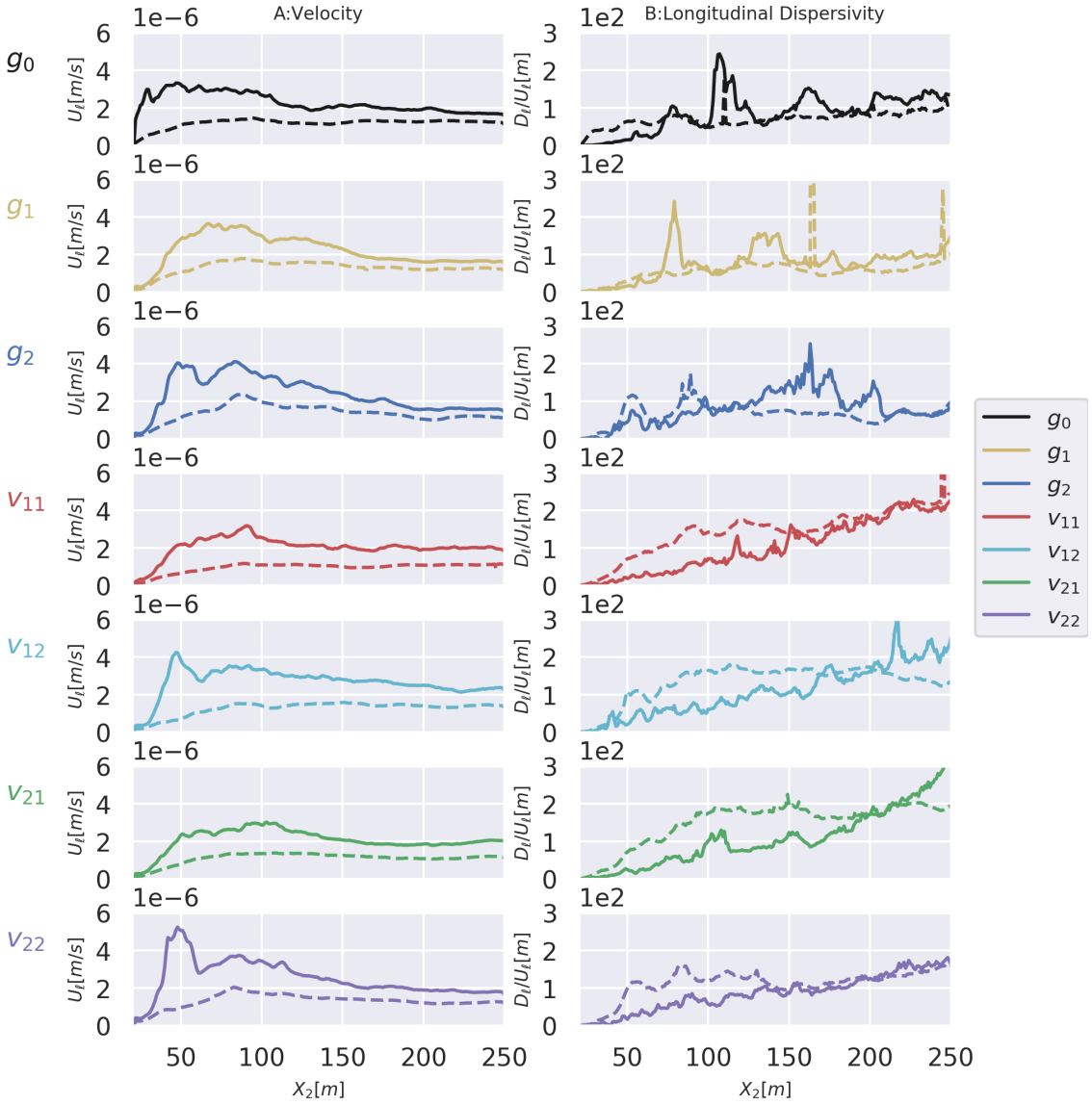

Figure 6.9: Macroscopic transport parameters derived from the travel-time distributions in observation planes perpendicular to the mean direction of flow. A: Effective (solid lines) and ensemble (dash lines) velocity; B: Effective (solid lines), and ensemble (dash lines) longitudinal dispersivity, D_ℓ/U_ℓ.

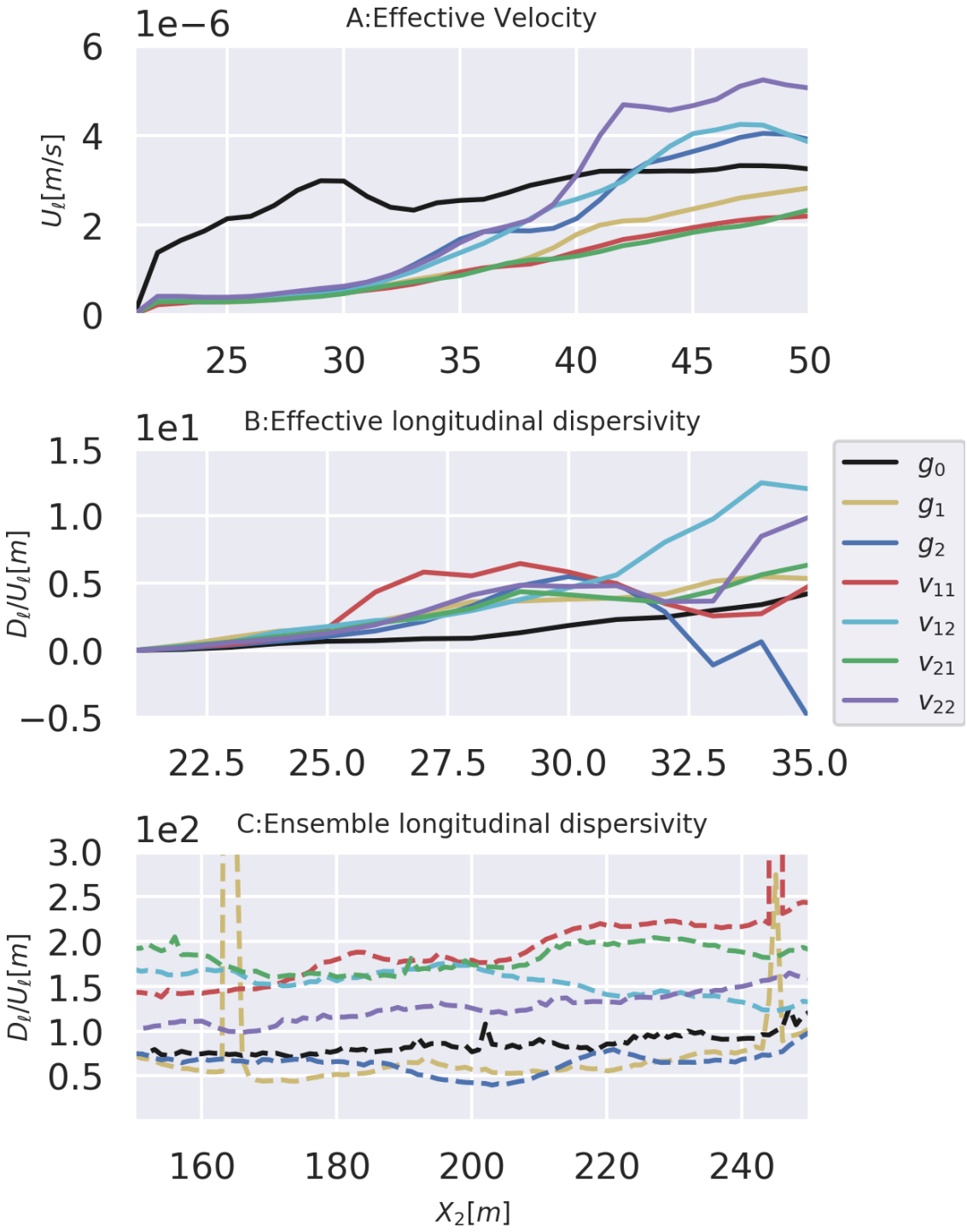

Figure 6.10: Comparison between macroscopic transport parameters among the different simulation models as function of travel distance. A: effective velocity $U_{\ell,eff}$; B: effective longitudinal dispersivity $D_{\ell,eff}/U_{\ell,eff}$; C: ensemble longitudinal dispersivity $D_{\ell,ens}/U_{\ell,ens}$

.

.

Figure 6.9 presents a comparison of the individual transport parameters among the different simulation models. All conditional models show the effect of the low-conductivity zone downstream of the injection point illustrated in Figure 6.5. As indicated by Figure 6.9A, the effective velocities remain low in the first ~ 10 meters travel distance downstream of the injection point in the conditional simulations, whereas the unconditional simulations don't show this behavior. After conditioning the K-field to the DPIL data, the conductivity along the plume trajectory has increased at $x_2 \sim 40\,[m]$, causing higher effective velocities $U_{\ell,eff}$ in this region in the models g_2, v_{12}, and v_{22} (see Figures 6.10A), with the strongest effects being seen in model v_{22}, which is consistent with the high probability mass of smaller log-travel times for v_{22} at $x_2 = 40\,[m]$ in Figure 6.7A.

Conditioning also affects solute dispersion. Directly downstream of the injection point, a strong deviation between the ensemble and effective dispersivities is observed in the unconditional simulations (g_0). In some unconditional realizations, this zone exhibits high K values and in others low ones, which causes a high uncertainty in the mean travel time of plumes in individual realizations. This uncertainty is manifested in a rapid increase of the ensemble dispersivity $D_{\ell,ens}/U_{\ell,ens}$. Since all conditional simulations agree on this zone exhibiting low K values (see Figure 6.10A), the early longitudinal displacement is fairly uniform among the different conditional realizations, causing smaller ensemble dispersivities.

As seen in Figure 6.9B, the non-multi-Gaussian models (v_{11}-v_{22}) yield a strong difference between the ensemble and effective dispersivities that persists at least until $x_2 \sim 150[m]$, whereas these two dispersivities differ less in the conditional multi-Gaussian models (g_1, g_2). Figure 6.10C highlights the differences in the ensemble dispersivity among the different models between $x_2 = 150\,[m]$ and $x_2 = 250\,[m]$. Quite clearly, the non-multi-Gaussian model results in higher ensemble dispersivities, indicating a larger uncertainty in the mean transport of the plume. The effects are higher in the models that only account for the flowmeter data (v_{11} and v_{21}). This is related to the vertical structure of the mean velocity (see Figure 6.5A) and the higher variance of log-conductivity in the top layer of the domain (see Figure 6.5B). In some realizations, the tracer particles exclusively move through the high-velocity top layer, whereas in others they also experience lower velocities. As the multi-Gaussian fields don't show persistent shear-flow, there can also be no uncertainty about the position of the trajectories concerning the vertical mean velocity profile.

6.2.3 Simulations of the MADE Tracer Test

The tritium distribution of four discrete points in time at 27 days, 132 days, 224 days, and 328 days since injection in the longitudinal (x_2) direction are compared between the

observations during the MADE2 tracer test (Boggs *et al.*, 1993) and the simulation results of the presented various geostatistical models.

Particles are injected at 20 points representing the five injection locations and 500 particles are injected at each point mimicking the field conditions. An alternative approach to simulate the gravity-feed injection did not yield better results, but instead the spurious movement of the injected mass in the upstream direction. The measured tritium concentrations are integrated in the vertical direction (x_3), classified into 25 bins with a constant class width of 10 [m] and normalized using the total recovered tritium mass. This approximated normalized mass distribution is compared with the simulated normalized mass distribution (Figure 6.11). The dense K observation network directly downstream of the injection locations determines the K-field in this zone also indicated by a small uncertainty in this zone (right column of Figure 6.5). This well-defined K-field leads to small differences in the simulated tritium mass distribution between various conditional geostatistical models (Figure 6.11A-D up to $x_2 \approx 100\,[m]$. The numerical tracer tests based on all geostatistical models can reproduce the first peak after the injection location and the shape of the early-time tritium distribution better than the unconditional model (g_0).

The solute transport behavior at medium to large travel distances (Figure 6.11B and C), starting at $x_2 \approx 100\,[m]$, might include a second peak at $x_2 \approx 200\,[m]$, is more challenging to be reproduced. However, geostatistical models that do include asymmetry (v_{11}, v_{12}, v_{21}, and v_{22}) match the observations at 132 days and at 224 days and associated large distances as well. Model v_{12} matches the long tail up to $x_2 = 250\,[m]$ (including the second peak) best at 224 days. The underestimation of the second peak in the numerical tracer tests might be attributed to a lack of important local features in the K-field due to the smaller number of K observations at $x_2 \approx 180\,[m]$ or to the steady hydraulic conditions in the numerical models whereas transient conditions might have occurred in the field (Llopis-Albert and Capilla, 2009).

The Cramér-von Mises criterion ω^2 (Anderson, 1962)

$$\omega^2 = \int_{.\infty}^{+\infty} [F_N(x) - F(X)]^2 dF(x) \tag{6.4}$$

is used to compare the observed ($F(x)$) and simulated ($F_N(x)$) tritium distribution in the longitudinal direction at two scales (the concentration-space and the log-concentration space). The smaller ω^2, the larger the similarity between two distributions. In the concentration-space (Figure 6.11E), ω^2 is mainly influenced by the high tritium mass density directly downstream of the injection location which dominates the global shape of the mass distribution. Hence, the conditional models ($g_1 - v_{22}$) match better the observations in that zone than the unconditional model g_0.

In the log-concentration space (Figures 6.11 A, B, C, D, F), small values can be compared better between observation and simulations. This focuses on the transport behavior at the zone farther away from the injection location. In that zone, the geostatistical models that include the measure of asymmetry A performed best. In this zone, the K observation network density is sparse, but the copula asymmetry provides at least some information globally, even when no observations are available. In this zone, the model with only flowmeter values (g_1) and/or the order of the DPIL values, exhibits the worst results. Even more, not only the asymmetry but also the order information from the DPIL measurements are meaningful to understand the solute transport behavior at the MADE site: an improved result including the order of the DPIL values can be found (Figure 6.11F) starting at day 132. Similarly, the velocity peaks at $x_2 \approx 50\,[m]$ (Figure 6.9) occurred when the DPIL values were included.

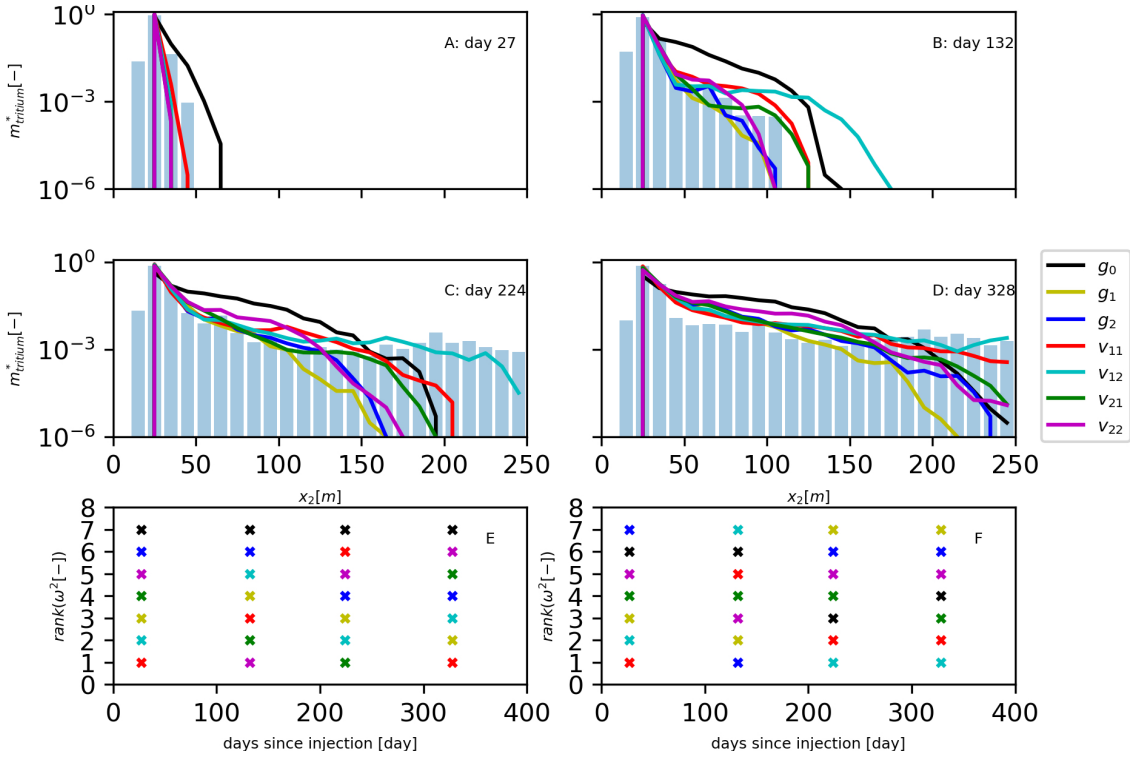

Figure 6.11: Observed (bar plot) and simulated normalized mass distributions of different geostatistical models of A) 27 days; B) 132 days; C) 224 days; D) 328 days since the injection and E) the Cramér-von Mises criterion of different models and F) the rank of the Cramér-von Mises criterion of different models in log scale.

6.3 Summary and Conclusion of this Chapter

In this chapter, the phase-annealing method (Chapter 3) is applied to the MADE site, where local hydraulic-conductivity estimates by the vertical flowmeter and direct-push injection-logging (DPIL) profiles exist, which are inconsistent in their absolute values but consistent in their rank-correlation function (Chapter 5). The data also show a strong depth dependence of the asymmetry function, indicating better connected high-K values at the top, indifferent asymmetry in the middle, and slightly better connected low-K values at the bottom. The effects of including the information from the flowmeter data only versus the flowmeter and DPIL datasets jointly, as well as accounting for the depth-dependent asymmetry of the K copulas are systematically tested. All models shared the same variogram and marginal distribution of K, but they differed in local features of the K-field and the transport characteristics revealed by particle-tracking random-walk simulations (Chapter 4) of a point-injection experiment.

Conditioning on point data alone has a significant effect on the resolved hydraulic-conductivity fields if the observation points are close-by to each other. At the MADE site, this is the case directly downstream of the injection location of the MADE-2 experiment. With a sufficiently dense observation network, the impact of the exact geostatistical model vanishes, because the conductivity field is largely controlled by the measured values at the conditioning points. In all conditional models tested, a low-K zone directly downstream of the injection exerts a major control on the short-distance plume behavior is confirmed. The DPIL data revealed an additional high-K zone shortly downstream of the low-K zone already detected by Rehfeldt *et al.* (1992). These features are particularly important for the interpretation of the MADE-2 experiment as they are close to the injection point.

Accounting for non-Gaussian spatial dependence becomes more important when the density of observation points is smaller as is the case over the largest part of the MADE site. In multi-Gaussian geostatistical methods, the impact of high- and low-K measurements on interpolated values between observation points decreases identically with the correlation function. At the MADE site, this implies conductivity fields that are hardly affected by conditioning in the areas between the clusters of observation points when multi-Gaussian conditioning methods are applied. The high positive asymmetry in the top layer of the MADE site, however, triggers a persistent high-K zone at the top upon depth-dependent non-multi-Gaussian conditioning. This layer causes a preferential flow layer at the top: Once the tracer plume has passed through the initial low-K zone, it is diverted into the top layer and stays therein. The multi-Gaussian model predicted neither this layer nor its effect on macroscopic solute transport. The pseudo-stratification in the non-multi-Gaussian models also caused larger ensemble dispersion coefficients and more persistent differences between ensemble and effective dispersion.

This chapter has shown that large-scale, three-dimensional conditional, non-multi-Gaussian simulations are feasible and can be applied to the benchmark of the highly heterogeneous MADE site, revealing persistent features in the K-field and transport characteristics that would remain unresolved with common geostatistical methods. third-order measures of non-Gaussianity (such as depth-dependent asymmetry) can improve the understanding of the spatial structure of K and its effects on flow and transport.

Chapter 7

Conclusions and Outlook

In this thesis, the copula-based three-dimensional non-multi-Gaussian conditional simulation shows the ability to build a field-scale data-driven stochastic model of the hydraulic conductivity for the study of solute transport at a real-world heterogeneous field site, namely the MAcroDispersion Experiment site. Three modeling and simulation techniques are combined, i.e., Copula-based geostatistical methods (Chapter 2) for the data analysis and modeling; Phase-annealing method with FFT-Asymmetry for non-multi-Gaussian K fields simulation (Chapter 3); GPU-accelerated particle-tracking random-walk method to evaluate the solute transport behavior on the simulated K fields (Chapter 4).

Copula-based geostatistical methods (Chapter 2) are capable to model a spatial dependence structure of K fields independent on the marginal distribution and are not influenced by a monotonic transformation. After removing the influences from the marginal distribution and the data transformation, e.g., the widely used log-transformation in hydrogeology, the underlying spatial dependence structure of DPIL and flowmeter dataset at the MADE site is discussed in detail in the copula space. The similarity between the copula-based univariate and bivariate measures further confirms that the flowmeter and the DPIL dataset present the same spatial dependence structure at the MADE site (Section 5.3). Although differences can be found between their marginal distributions. However, a difference between the two datasets has been found at certain locations (Figures 5.5 and 5.6). This could be a possible start point for future fieldwork to understand the K distribution at the MADE site and the relationship between the two datasets.

A non-multi-Gaussian spatial dependence structure with a different arrangement of two-point and n-point spatial dependence is found at the MADE site using the empirical bivariate copula asymmetry (Section 5.3) and the theoretical Gauss- and v-copula model (Section 5.3.2). The modeling of the spatial dependence structures of the DPIL and the flowmeter dataset has been improved by including the censorship threshold in the copula parameter estimation (Section 5.3.3). Further investigations like the definition of new theoretical copula models or a two-side censorship threshold and its inclusion in the

geostatistical simulation would be interesting topics for future work.

The influence of the degree of non-Gaussianity on solute transport has been studied using two types of marginal distribution and three different types of degrees of Gaussianity (Section 4.2). The empirical bivariate copula density of the longitudinal particle velocity proves the influence of the underlying spatial dependence structure. However, the corresponding influence on the macrodispersivities between a multi-Gaussian and non-multi-Gaussian scenario is small in the test scenarios with a small $\sigma^2(\ln(K))$. In contrast, a significant large influence can be found in the scenarios with a large $\sigma^2(\ln(K))$. This proves that the degree of the non-Gaussianity can not be neglected for a system with a large $\sigma^2(\ln(K))$. When this degree of non-Gaussianity coincides with the preferential flow paths, more complex behavior can be found.

A potential extension of copula-based geostatistics in the future is to model the bivariate copula density of the longitudinal particle velocity. The transition of the longitudinal velocity of the injected particles can be assumed as a n-step spatial Markov process and simulated using the bivariate copula density. Critical points of this method are how to model the particle behavior after the injection point and how to model a non-stationary process.

The influence of the underlying degree of non-Gaussianity and local features on solute transport is studied for the MADE site (Chapter 6). The variogram, point values of the flowmeter dataset, the order of the point values of the DPIL dataset, and the depth-dependent asymmetry are combined in a single PA simulation to extract more information from the K observations. The local feature after the injection point generates an uplifting path and pushes the particles to an over-averaging high-K zone. These together cause a long tail particle distribution and additional macrodispersivity.

The copula asymmetry estimates additional information from the existing observations and provides information on the relative location of the K values of different quantiles in rank space and the relative location between the injection point and the important K-zones away from Gaussian. When the investigated process is strongly influenced by these heterogeneous structures, the included asymmetry can help to improve the model accuracy. An interesting extension in the future is estimating the three-dimensional asymmetry from the geophysical data.

The updating of the asymmetry in annealing iteration occupies most of the computational time, although it is simplified in this thesis. Possible improvements to address this issue are:

1. the development of a wavelet-based phase-annealing method. Therefore in this method, the values are updated locally during the annealing iterations. The asymmetry can be updated using Equation 3.2;

2. investigating the relationship between asymmetry in the space domain and the phase angle distributions in the Fourier domain to determine the most important wave number or the pattern of phase distributions.

Phase-annealing (Chapter 6) has been shown as a powerful tool for the three-dimensional conditional stochastic simulation of random fields with variogram. One PA-simulation can combine different types of information and uncertainty. It can be further used to simulate high-resolution porous media for the application to environmental, material science, or industrial problems. A particularly interesting question for further studies is the numerical properties of PA-simulation while including different terms in the objective function or a combination of the existing multi-objective optimization techniques in PA-simulation.

Appendix A

FFT Representation of the Asymmetry

The cross-covariance between two random filed $f(\mathbf{x})$ and $g(\mathbf{x})$ is:

$$Cov(\mathbf{h}) = \int_{-\infty}^{+\infty} f(x)g(x+h). \tag{A.1}$$

Equation A.1 can be described using the Fourier transform:

$$Cov(\mathbf{h}) = \mathbf{TF}^{-1}(\mathbf{TF}(f(u))\overline{\mathbf{TF}(\mathbf{g}(\mathbf{u}))}), \tag{A.2}$$

in which **TF** is the Fourier transform, $\mathbf{TF}^{-1}$ is the inverse Fourier transform, and $\overline{(\cdot)}$ is the complex conjugate. Let $f = F_z(z(\mathbf{x}))$, $g = F_z(z(\mathbf{x}+\mathbf{h}))$, and indicator matrices $I_f = I(z(\mathbf{x}))$, $I_g = I(z(\mathbf{x}+\mathbf{h}))$, then the asymmetry A in Equations2.27 and 2.28 can be rewritten as:

$$\begin{aligned}
&\hat{A}(f,g) \\
&\quad = \left(\left(f-\frac{1}{2}\right)^2\left(g-\frac{1}{2}\right)\right) + \left(\left(f-\frac{1}{2}\right)\left(g-\frac{1}{2}\right)^2\right) \\
&\quad = \left(f^2g - \frac{1}{2}f^2 + \frac{1}{4}g - \frac{1}{8} - fg + \frac{1}{2}f\right) \\
&\quad + \left(g^2f - \frac{1}{2}g^2 + \frac{1}{4}f - \frac{1}{8} - gf + \frac{1}{2}g\right) \\
&\quad = \left(g^2f + f^2g - \frac{1}{2}g^2 - \frac{1}{2}f^2 - 2gf + \frac{3}{4}g + \frac{3}{4}f - \frac{1}{4}\right) \\
&\quad = \left(g^2f + f^2g - \frac{1}{2}g^2I_f - \frac{1}{2}f^2I_g - 2gf + \frac{3}{4}gI_f + \frac{3}{4}fI_g - \frac{1}{4}I_fI_g\right)
\end{aligned} \tag{A.3}$$

and

$$
\begin{aligned}
&A\,(f,g))\\
&\quad=\frac{\sum\left(g^2f+f^2g-\frac{1}{2}g^2I_f-\frac{1}{2}f^2I_g-2gf+\frac{3}{4}gI_f+\frac{3}{4}fI_g-\frac{1}{4}I_fI_g\right)}{\sum I_fI_g}\\
&\quad=\frac{\mathbf{TF}^{-1}\left(\overline{G_2}F_1+\overline{G_1}F_2-\frac{1}{2}\overline{G_2}F_{I_f}-\frac{1}{2}\overline{F_{I_g}}F_2-2\overline{G_1}F_1+\frac{3}{4}\overline{G_1}F_{I_f}+\frac{3}{4}\overline{F_{I_g}}F_1-\frac{1}{4}\overline{F_{I_g}}F_{I_f}\right)}{\mathbf{TF}^{-1}(\overline{F_{I_g}}F_{I_f})},
\end{aligned}
\tag{A.4}
$$

in which $F_1=\mathbf{TF}(f)$, $F_2=\mathbf{TF}(f^2)$, $G_1=\mathbf{TF}(g)$, $G_2=\mathbf{TF}(g^2)$, $F_{I_f}=\mathbf{TF}(I_f)$ and $F_{I_g}=\mathbf{TF}(I_g)$.

The asymmetry in Equations 2.27 and 2.28 is a auto-asymmetry in this thesis. Therefore $F_{I_f}=F_{I_g}$, $F_1=G_1$, $F_2=G_2$. So, Equation A.4 can be rewritten as:

$$
\begin{aligned}
&A\\
&=\frac{\mathbf{TF}^{-1}\left(\overline{F_2}F_1+\overline{F_1}F_2-\frac{1}{2}\overline{F_2}F_I-\frac{1}{2}\overline{F_I}F_2-2\overline{F_1}F_1+\frac{3}{4}\overline{F_1}F_I+\frac{3}{4}\overline{F_I}F_1-\frac{1}{4}\overline{F_I}F_I\right)}{\mathbf{TF}^{-1}(\overline{F_I}F_I)}.
\end{aligned}
\tag{A.5}
$$

Equation A.5 can be further simplified to reduce the computational cost. Let $F_{1,k}=a_1+b_1i$, $F_{2,k}=a_2+b_2i$ and $F_{I,k}=a_0+b_0i$ are the $k-$th term in the Fourier space. The numerator part of the inverse FT of Equation A.5 can be rewritten as:

$$
\begin{aligned}
B=&\,(a_2a_1+a_2b_1i-b_2a_1i+b_2b_1)\\
&+(a_1a_2+a_1b_2i-b_1a_2i+b_1b_2)\\
&-\frac{1}{2}(a_2a_0+a_2b_0i-b_2a_0i+b_2b_0)\\
&-\frac{1}{2}(a_0a_2+a_0b_2i-b_0a_2i+b_0b_2)\\
&-(2a_1^2+2b_1^2)\\
&+\frac{3}{4}(a_1a_0+a_1b_0i-b_1a_0i+b_1b_0)\\
&+\frac{3}{4}(a_0a_1+a_0b_1i-b_0a_1i+b_0b_1)\\
&-\frac{1}{4}(a_0^2+b_0^2)\\
&=2(a_1a_2+b_1b_2)-(a_0a_2+b_0b_2)-2(a_1^2+b_1^2)+\frac{3}{2}(a_0a_1+b_0b_1)-\frac{1}{4}(a_0^2+b_0^2).
\end{aligned}
\tag{A.6}
$$

The first term $2(a_1a_2+b_1b_2)$ in Equation A.6 is the same as the FFT representation of

$$\hat{A}\left(F_z(z(\mathbf{x_i})),F_z(z(\mathbf{x_j}))\right) = F_z(z(\mathbf{x_i}))^2(F_z(z(\mathbf{x_j}))+F_z(z(\mathbf{x_i}))(F_z(z(\mathbf{x_j}))^2, \tag{A.7}$$

so the additional terms are introduced by the central part -0.5. The asymmetry A can be calculated by

$$A = \frac{\mathbf{TF}^{-1}\left(2(a_1a_2+b_1b_2)-(a_0a_2+b_0b_2)-2(a_1^2+b_1^2)+\frac{3}{2}(a_0a_1+b_0b_1)-\frac{1}{4}(a_0^2+b_0^2)\right)}{\mathbf{TF}^{-1}(a_0^2+b_0^2)}, \tag{A.8}$$

in which a_0, b_0 are constants and can be calculated outside of the annealing iterations.

A further simplification can be archived using $\tilde{f} = f-0.5$, $\tilde{g} = g-0.5$, $\tilde{F_1} = \mathbf{TF}(\tilde{f})$, $\tilde{G_1} = \mathbf{TF}(\tilde{g})$, $\tilde{F_2} = \mathbf{TF}(\tilde{f}^2)$ and $\tilde{G_2} = \mathbf{TF}(\tilde{g}^2)$, Equation A.3, A.4 and A.8 can be rewritten as

$$\hat{A}\left(\tilde{f},\tilde{g}\right) = \tilde{f}^2\cdot\tilde{g}+\tilde{f}\cdot\tilde{g}^2 \tag{A.9}$$

and

$$\begin{aligned} A\left(\tilde{f},\tilde{g}\right)) &= \frac{\sum\left(\tilde{g}^2\tilde{f}+\tilde{f}^2\tilde{g}\right)}{\sum I_f I_g} \\ &= \frac{\mathbf{TF}^{-1}\left(\overline{\tilde{G_2}}\tilde{F_1}+\overline{\tilde{G_1}}\tilde{F_2}\right)}{\mathbf{TF}^{-1}(\overline{F_{I_g}}F_{I_f})} \\ &= \frac{\mathbf{TF}^{-1}\left(2(\tilde{a_1}\tilde{a_2}+\tilde{b_1}\tilde{b_2})\right)}{\mathbf{TF}^{-1}(a_0^2+b_0^2)}, \end{aligned} \tag{A.10}$$

in which $\tilde{F_1} = \tilde{G_1} = \tilde{a_1}+\tilde{b_1}i$ and $\tilde{F_2} = \tilde{G_2} = \tilde{a_2}+\tilde{b_2}i$. The asymmetry A is equivalent to the cross-covariance of $F_z(z(\mathbf{x}))-0.5$ and $(F_z(z(\mathbf{x}+\mathbf{h}))-0.5)^2$. $A \stackrel{!}{=} 0$ for a perfect Gaussian dependence, in this case $(\tilde{a_1}\tilde{a_2}+\tilde{b_1}\tilde{b_2}) \stackrel{!}{=} 0$ and the cross-covariance of $F_z(z(\mathbf{x}))-0.5$ and $(F_z(z(\mathbf{x}+\mathbf{h}))-0.5)^2$ is zero.

Equation A.10 can be written as:

$$
\begin{aligned}
& A\left(\tilde{f},\tilde{g}\right)\right) \\
& \quad = \frac{\mathbf{TF}^{-1}\left(2(\tilde{a_1}\tilde{a_2}+\tilde{b_1}\tilde{b_2})\right)}{\mathbf{TF}^{-1}(a_0^2+b_0^2)} \\
& \quad = \frac{\mathbf{TF}^{-1}\left(2<\tilde{F_1},\tilde{F_2}>\right)}{\mathbf{TF}^{-1}(a_0^2+b_0^2)} \\
& \quad = \frac{\mathbf{TF}^{-1}\left(2<\tilde{F_1},\tilde{F_1}*\tilde{F_1}>\right)}{\mathbf{TF}^{-1}(a_0^2+b_0^2)} \\
& \quad = \frac{\mathbf{TF}^{-1}\left(2<\tilde{F_1},R_{xx}(\tilde{F_1})>\right)}{\mathbf{TF}^{-1}(a_0^2+b_0^2)},
\end{aligned}
\tag{A.11}
$$

in which $<\cdot,\cdot>$ is the inner product, $*$ is the convolution operator and R_{xx} is the auto-correlation function. Equation A.11 connects the structure of asymmetry and the structure of the Fourier coefficients. According to Equation A.11, asymmetry is a function of $\tilde{F_1} = FT(F_Z(z(\mathbf{x}))-0.5)$ and $R_{xx}(\tilde{F_1}) = R_{xx}(FT(F_Z(z(\mathbf{x}))-0.5))$. In phase-annealing, $\tilde{F_1}$ is changed only on the locations with a phase randomization and R_{xx} can be updated. In the case of a Gaussian random field, there is $<\tilde{F_1},R_{xx}(\tilde{F_1})>=0$. From the geometric point, this means $\tilde{F_1}$ is orthogonal to $R_{xx}(\tilde{F_1})$.

In the annealing iteration, there are $2\times\mathbf{FT}(\cdot)$, $2\times\mathbf{FT^{-1}}(\cdot)$ and $1\times\mathbf{FTShift}(\cdot)$. Because FTshift function is not implemented in pyFFTW, there are two different methods to reduce its computational cost. a) The asymmetry values are taken directly from the asymmetry map from Equation A.8 without doing the FTshift. 2) A FTshift function swaps values between blocks in 2D and cubes in 3D. A thread-safe Algorithm 6 (Abdellah, 2014) is implemented by Cython to perform a multithread FTshift.

Algorithm 6: Thread-safe FTshift for a three-dimensional even size random field.

Input: RF $\mathbf{g}$ before FT shift with a domain size $\mathbf{d} = [2 \cdot N_1, 2 \cdot N_2, 2 \cdot N_3]$
Result: Shifted RF
initialization;
$\mathbf{d}_2 \leftarrow \mathbf{d}/2$;
for *i in* $[0, \mathbf{d}_2[0])$ **do**
  for *j in* $[0, \mathbf{d}_2[1])$ **do**
    for *k in* $[0, \mathbf{d}_2[2])$ **do**
      $\mathbf{g}[i,j,k], \mathbf{g}[i+\mathbf{d}_2[0], j+\mathbf{d}_2[1], k+\mathbf{d}_2[2]] \leftrightarrow$
      $\mathbf{g}[i+\mathbf{d}_2[0], j+\mathbf{d}_2[1], k+\mathbf{d}_2[2]], \mathbf{g}[i,j,k]$;
      $\mathbf{g}[i+\mathbf{d}_2[0], j, k], \mathbf{g}[i, j+\mathbf{d}_2[1], k+\mathbf{d}_2[2]] \leftrightarrow$
      $\mathbf{g}[i, j+\mathbf{d}_2[1], k+\mathbf{d}_2[2]], \mathbf{g}[i+\mathbf{d}_2[0], j, k]$;
      $\mathbf{g}[i, j+\mathbf{d}_2[1], k], \mathbf{g}[i+\mathbf{d}_2[0], j, k+\mathbf{d}_2[2]] \leftrightarrow$
      $\mathbf{g}[i+\mathbf{d}_2[0], j, k+\mathbf{d}_2[2]], \mathbf{g}[i, j+\mathbf{d}_2[1], k]$;
      $\mathbf{g}[i, j, k+\mathbf{d}_2[2]], \mathbf{g}[i+\mathbf{d}_2[0], j+\mathbf{d}_2[1], k] \leftrightarrow$
      $\mathbf{g}[i+\mathbf{d}_2[0], j+\mathbf{d}_2[1], k], \mathbf{g}[i, j, k+\mathbf{d}_2[2]]$;
    end
  end
end

Appendix B

Performance of Python Scientific Libraries

The main performance bottleneck of the PA with asymmetry is the updating of the asymmetry in each PA iteration, which contains a forward/inverse Fourier Transformation and matrix operation. The computational costs of different Python libraries are compared and the most efficient one could be selected according to the available computational resources to get the best computational performance.

B.1 Configuration of the Test Machine

Table B.1: Configurations of the Test Machine.

	TM1	TM2
CPU	Intel i7-4790	Intel Xeon E5-2660 v3
Frequency	3.60 GHz	2.60 Ghz
Threads	8 threads	20 threads
Memory	16 GB	256 GB
GPU	Nvidia GTX970	
GPU Memory	4GB	
OS	Windows 7 Professional	CentOS Linux 7
OS Architecture	x86 64-bit	x86 64-bit

Test machines with two different configurations are used. One is local PC (TM1) and one is on a high-performance cluster (TM2) (Table B.1).

B.2 Fourier Transformation

The performance of the FFT operation (forward FFT and inverse FFT) depends on the implementation of the library and the hardware: numpy, scipy, and pyFFTw on CPU; cupy on GPU, number of the thread: single thread or multithread, the data type: 32- or 64-bit float number. According to these controlling parameters, the following scenarios are tested:

1. np1: using numpy to generate the 64-bit input random matrixes and FFT operation;
2. np2: using numpy to generate the 32-bit input random matrixes and FFT operation;
3. np3: using numpy to generate the 32-bit input random matrixes and real value FFT operation;
4. scipy1: using numpy to generate the 64-bit input random matrixes and scipy to perform the FFT operation;
5. scipy2: using numpy to generate the 32-bit input random matrixes and scipy to perform the FFT operation;
6. scipy3: using numpy to generate the 32-bit input random matrixes and scipy to perform the real value FFT operation;
7. cp1: using cupy to generate the 64-bit input random matrixes and FFT operation;
8. cp2: using cupy to generate the 32-bit input random matrixes and FFT operation;
9. cp3: using cupy to generate the 32-bit input random matrixes and real value FFT operation;
10. cp4: using numpy to generate the 64-bit input random matrixes and cupy to perform the FFT operation;
11. cp5: using numpy to generate the 32-bit input random matrixes and cupy to perform the FFT operation;
12. cp6: using numpy to generate the 32-bit input random matrixes and cupy to perform

the real value FFT operation;

13. fftw1: using numpy to generate the 64-bit input random matrixes and pyfftw to perform the FFT operation with 8-thread;
14. fftw2: using numpy to generate the 64-bit input random matrixes and pyfftw to perform the FFT operation with 6-thread;
15. fftw3: using numpy to generate the 64-bit input random matrixes and pyfftw to perform the FFT operation with 4-thread;
16. fftw4: using numpy to generate the 32-bit input random matrixes and pyfftw to perform the FFT operation with 8-thread;
17. fftw5: using numpy to generate the 32-bit input random matrixes and pyfftw to perform the FFT operation with 6-thread;
18. fftw6: using numpy to generate the 32-bit input random matrixes and pyfftw to perform the FFT operation with 4-thread;
19. fftw7: using numpy to generate the 64-bit input random matrixes and pyfftw to perform the real value FFT operation with 8-thread;
20. fftw8: using numpy to generate the 64-bit input random matrixes and pyfftw to perform the real value FFT operation with 6-thread;
21. fftw9: using numpy to generate the 64-bit input random matrixes and pyfftw to perform the real value FFT operation with 4-thread;
22. fftw10: using numpy to generate the 32-bit input random matrixes and pyfftw to perform the real value FFT operation with 8-thread;
23. fftw11: using numpy to generate the 32-bit input random matrixes and pyfftw to perform the real value FFT operation with 6-thread;
24. fftw12: using numpy to generate the 32-bit input random matrixes and pyfftw to perform the real value FFT operation with 4-thread.

Tables B.2 and B.3 show the performance test results on the local PC (TM1) and on the cluster (TM2). The first scenario np1 on TM1 is used as the baseline scenario. So, a value larger than one means better performance than the baseline scenario np1.

The scenarios of cupy (cp1-cp3) show that the FFT operation on GPU is faster than on CPU, especially for 32-bit data (cp2 and cp3). Scenario with PyFFTW (fftw1-fftw12)

provides good performance of real value FFT on 32-bit data with 8-thread (fftw10-fftw12). Additionally, the performance of PyFFTW on the cluster (Table B.3) is better than on the local PC (Table B.2).

Table B.2: Performance test of FFT Transformation on TM2.

d	100^3	128^3	200^3	250^3	256^3
np1	1.00	1.00	1.00	1.00	1.00
np2	1.00	1.02	1.00	1.01	1.00
np3	1.54	1.64	1.53	1.56	1.83
scipy1	1.40	1.33	1.40	1.42	1.10
scipy2	1.69	1.37	1.71	1.64	1.40
scipy3	2.30	2.38	2.26	2.30	2.84
cp1	19.56	6.30	256.71	6.22	13.76
cp2	32.25	84.55	282.31	125.44	753.20
cp3	12.51	31.31	104.64	222.19	288.87
cp4	5.88	8.17	8.39	9.29	11.09
cp5	6.92	9.54	9.93	10.72	12.79
cp6	6.87	9.50	9.89	10.67	12.79
fftw1	4.01	3.08	3.82	3.90	2.78
fftw2	3.00	2.42	3.14	3.73	2.42
fftw3	4.01	2.85	3.51	3.62	2.35
fftw4	4.54	3.70	4.56	4.47	3.46
fftw5	3.98	2.98	3.09	3.11	3.16
fftw6	4.64	3.43	4.24	4.05	2.87
fftw7	5.43	5.76	5.35	5.62	6.50
fftw8	4.75	5.25	5.00	5.02	6.14
fftw9	5.46	5.78	5.10	5.31	6.06
fftw10	6.31	7.07	6.41	6.52	7.97
fftw11	5.78	6.42	5.75	5.93	7.16
fftw12	6.44	7.13	5.95	6.06	7.48

Table B.3: Performance test of FFT Transformation on the cluster.

d	100^3	128^3	200^3	250^3	256^3
np1	0.80	0.71	0.65	0.65	0.74
np2	0.72	0.69	0.64	0.64	0.75
np3	1.36	1.39	1.08	1.08	1.30
fftw1	4.74	5.14	4.63	4.81	5.17
fftw2	3.14	3.42	4.42	4.61	3.67
fftw3	4.27	4.44	4.04	4.17	4.45
fftw4	5.78	6.50	5.56	5.73	6.25
fftw5	2.53	6.31	2.58	5.52	5.90
fftw6	5.35	5.95	5.06	5.14	5.24
fftw7	5.79	6.67	5.62	5.85	7.01
fftw8	5.21	6.42	5.06	5.32	6.40
fftw9	5.31	6.04	5.00	5.22	6.28
fftw10	6.76	7.91	7.01	6.80	8.16
fftw11	6.64	7.48	3.95	6.41	8.01
fftw12	6.41	7.51	6.51	6.30	7.66

B.3 Matrix Operation

The performance of the matrix operation (+, -, *, and /) is compared between numpy, Cython, cupy on three-dimensional matrixes of different sizes:

1. np1: using numpy to generate random matrixes and perform the matrix operation;
2. cp1: using cupy to generate random matrixes and perform the matrix operation;
3. cp2: using numpy to generate random matrixes and cupy to perform the matrix operation;
4. cython1: using numpy to generate random matrixes and cython to perform the matrix operation with 6-thread;

5. cython2: using numpy to generate random matrixes and cython to perform the matrix operation with 8-thread.

Table B.4 shows the performance test results on the local PC (TM1) and on the cluster (TM2). The first scenario np1 on TM1 is used as the baseline scenario. So, a value larger than one means better performance than the baseline scenario np1.

The scenarios of cupy (cp1 and cp2) show that the matrix operation on GPU is faster than on CPU. This computational benefit is reduced when a large matrix is exchanged between GPU and CPU (cp2). Scenario with Cython (cython1 and cython2) provides good performance with 8-thread on the cluster (cython2).

Table B.4: Performance test of matrix operation.

d	50^3	100^3	150^3	200^3	250^3
On TM1					
np1	1.00	1.00	1.00	1.00	1.00
cp1	4.67	25.13	25.31	25.02	24.81
cp2	0.27	3.26	2.98	3.07	3.11
cython1	0.74	1.75	1.65	1.67	1.68
cython2	0.74	1.75	1.64	1.67	1.68
On TM2					
np1	0.73	2.58	1.74	1.01	1.02
cython1	0.43	3.03	3.50	3.74	6.72
cython2	0.45	3.12	3.66	3.79	5.89

List of Tables

List of Figures

List of Algorithms

Bibliography

Abdellah, M. (2014). CufftShift: High Performance CUDA-Accelerated FFT-Shift Library. In *Proceedings of the High Performance Computing Symposium*, HPC '14, San Diego, CA, USA. Society for Computer Simulation International.

Abebe, A., Guinot, V., and Solomatine, D. (2000). Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters. *Proc. 4th International Conference on Hydroinformatics, Iowa City, USA*, (July).

Adams, E. E. and Gelhar, L. W. (1992). Field study of dispersion in a heterogeneous aquifer: 2. Spatial moments analysis. *Water Resources Research*, **28**(12), 3293–3307.

Anderson, T. W. (1962). On the Distribution of the Two-Sample Cramer-von Mises Criterion. *The Annals of Mathematical Statistics*, **33**(3), 1148–1159.

Attinger, S., Dentz, M., Kinzelbach, H., and Kinzelbach, W. (1999). Temporal behaviour of a solute cloud in a chemically heterogeneous porous medium. *Journal of Fluid Mechanics*, **386**, 77–104.

Bárdossy, A. (2006). Copula-based geostatistical models for groundwater quality parameters. *Water Resources Research*, **42**(11).

Bárdossy, A. (2011). Interpolation of groundwater quality parameters with some values below the detection limit. *Hydrology and Earth System Sciences*, **15**(9), 2763–2775.

Bárdossy, A. and Hörning, S. (2016). Gaussian and non-Gaussian inverse modeling of groundwater flow using copulas and random mixing. *Water Resources Research*, **52**(6), 4504–4526.

Bárdossy, A. and Hörning, S. (2017). Process-Driven Direction-Dependent Asymmetry: Identification and Quantification of Directional Dependence in Spatial Fields. *Mathematical Geosciences*, **49**(7), 871–891.

Bárdossy, A. and Li, J. (2008). Geostatistical interpolation using copulas. *Water Resources Research*, **44**(7).

Bárdossy, A. and Pegram, G. G. (2009). Copula based multisite model for daily precipitation simulation. *Hydrology and Earth System Sciences*, **13**(12), 2299–2314.

Barlebo, H. C., Hill, M. C., and Rosbjerg, D. (2004). Investigating the Macrodispersion Experiment (MADE) site in Columbus, Mississippi, using a three-dimensional inverse flow and transport model. *Water Resources Research*, **40**(4).

Benson, D. A., Schumer, R., Meerschaert, M. M., and Wheatcraft, S. W. (2001). Fractional Dispersion, Lévy Motion, and the MADE Tracer Tests. *Transport in Porous Media*, **42**(1), 211–240.

Berkowitz, B. and Scher, H. (1998). Theory of anomalous chemical transport in random fracture networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, **57**(5), 5858–5869.

Berkowitz, B., Cortis, A., Dentz, M., and Scher, H. (2006). Modeling Non-fickian transport in geological formations as a continuous time random walk. *Reviews of Geophysics*, **44**(2), RG2003.

Bianchi, M. and Pedretti, D. (2017). Geological entropy and solute transport in heterogeneous porous media. *Water Resources Research*, **53**(6), 4691–4708.

Bianchi, M. and Zheng, C. (2016). A lithofacies approach for modeling non-Fickian solute transport in a heterogeneous alluvial aquifer. *Water Resources Research*, **52**(1), 552–565.

Bianchi, M., Zheng, C., Wilson, C., Tick, G. R., Liu, G., and Gorelick, S. M. (2011). Spatial connectivity in a highly heterogeneous aquifer: From cores to preferential flow paths. *Water Resources Research*, **47**(5).

Boggs, J. M., Young, S. C., Beard, L. M., Gelhar, L. W., Rehfeldt, K. R., and Adams, E. E. (1992). Field study of dispersion in a heterogeneous aquifer: 1. Overview and site description. *Water Resources Research*, **28**(12), 3281–3291.

Boggs, J. M., Beard, L. M., Long, S. E., McGee, M. P., MacIntyre, W. G., Antworth, C. P., and Stauffer, T. B. (1993). Database for the second macrodispersion experiment (MADE-2). Technical report, Electric Power Research Institute.

Bohling, G. C., Liu, G., Knobbe, S. J., Reboulet, E. C., Hyndman, D. W., Dietrich, P., and Butler, J. J. (2012). Geostatistical analysis of centimeter-scale hydraulic conductivity variations at the MADE site. *Water Resources Research*, **48**(2).

Bohling, G. C., Liu, G., Dietrich, P., and Butler, J. J. (2016). Reassessing the MADE direct-push hydraulic conductivity data using a revised calibration procedure. *Water Resources Research*, **52**(11), 8970–8985.

Boisvert, J. B. and Deutsch, C. V. (2011). Programs for kriging and sequential Gaussian simulation with locally varying anisotropy using non-Euclidean distances. *Computers and Geosciences*, **37**(4), 495–510.

Bowling, J. C., Zheng, C., Rodriguez, A. B., and Harry, D. L. (2006). Geophysical constraints on contaminant transport modeling in a heterogeneous fluvial aquifer. *Journal of Contaminant Hydrology*, **85**(1-2), 72–88.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**(3), 345–370.

Carle, S. F. (1999). T-PROGS: Transition Probability Geostatistical Software.

Chen, Y. C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics and Epidemiology*, **1**(1), 161–187.

Chilès, J. P. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty: Second Edition*. John Wiley & Sons.

Cirpka, O. A. (2002). Choice of dispersion coefficients in reactive transport calculations on smoothed fields. *Journal of Contaminant Hydrology*, **58**(3-4), 261–282.

Cirpka, O. A. and Kitanidis, P. K. (2000). Characterization of mixing and dilution in heterogeneous aquifers by means of local temporal moments. *Water Resources Research*, **36**(5), 1221–1236.

Cohn, T. A. (2005). Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data. *Water Resources Research*, **41**(7), 1–13.

Cvetkovic, V., Cheng, H., and Wen, X. H. (1996). Analysis of nonlinear effects on tracer migration in heterogeneous aquifers using Lagrangian travel time statistics. *Water Resources Research*, **32**(6), 1671–1680.

Cvetkovic, V., Fiori, A., and Dagan, G. (2014). Solute transport in aquifers of arbitrary variability: A time-domain random walk formulation. *Water Resources Research*, **50**(7), 5759–5773.

Dagan, G. (1982a). Stochastic modeling of groundwater flow by unconditional and conditional probabilities: 1. Conditional simulation and the direct problem. *Water Resources Research*, **18**(4), 813–833.

Dagan, G. (1982b). Stochastic modeling of groundwater flow by unconditional and conditional probabilities: 2. The solute transport. *Water Resources Research*, **18**(4), 835–848.

Dagan, G., Cvetkovic, V., and Shapiro, A. (1992). A solute flux approach to transport in heterogeneous formations: 1. The general framework. *Water Resources Research*, **28**(5), 1369–1376.

Darcy, H. (1856). *Les fontaines publiques de la ville de Dijon*. Paris: Victor Dalmont.

Davis, M. W. (1987). Production of conditional simulations via the LU triangular decomposition of the covariance matrix. *Mathematical Geology*, **19**(2), 91–98.

De Michele, C. and Salvadori, G. (2003). A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas. *Journal of Geophysical Research D: Atmospheres*, **108**(2), 1–11.

Deans, S. R. (1983). *The Radon Transform and Some of Its Applications*. Wiley.

Delhomme, J. P. (1978). Kriging in the hydrosciences. *Advances in Water Resources*, **1**(5), 251–266.

Delhomme, J. P. (1979). Spatial variability and uncertainty in groundwater flow parameters: A geostatistical approach. *Water Resources Research*, **15**(2), 269–280.

Dentz, M., Kinzelbach, H., Attinger, S., and Kinzelbach, W. (2000). Temporal behavior of a solute cloud in a heterogeneous porous medium 1. Point-like injection. *Water Resources Research*, **36**(12), 3591–3604.

Dentz, M., Comolli, A., Hakoun, V., and Hidalgo, J. J. (2020). Transport Upscaling in Highly Heterogeneous Aquifers and the Prediction of Tracer Dispersion at the MADE Site. *Geophysical Research Letters*, **47**(22).

Deutsch, C. V. and Cockerham, P. W. (1994). Practical considerations in the application of simulated annealing to stochastic simulation. *Mathematical Geology*, **26**(1), 67–82.

Dietrich, C. R. and Newsam, G. N. (1993). A fast and exact method for multidimensional gaussian stochastic simulations. *Water Resources Research*, **29**(8), 2861–2869.

Dietrich, C. R. and Newsam, G. N. (1997). Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal of Scientific Computing*, **18**(4), 1088–1107.

Dogan, M., Van Dam, R. L., Liu, G., Meerschaert, M. M., Butler, J. J., Bohling, G. C., Benson, D. A., and Hyndman, D. W. (2014). Predicting flow and transport in highly heterogeneous alluvial aquifers. *Geophysical Research Letters*, **41**(21), 7560–7565.

Dünser, S. and Meyer, D. W. (2016). Predicting field-scale dispersion under realistic conditions with the polar Markovian velocity process model. *Advances in Water Resources*, **92**, 271–283.

Edery, Y., Guadagnini, A., Scher, H., and Berkowitz, B. (2014). Origins of anomalous transport in heterogeneous media: Structural and dynamic controls. *Water Resources Research*, **50**(2), 1490–1505.

Embrechts, P., Lindskog, F., and Mcneil, A. (2003). Modelling Dependence with Copulas and Applications to Risk Management. Technical report, Department of Mathematics ETHZ.

Erhardt, T. M., Czado, C., and Schepsmeier, U. (2015a). R-vine models for spatial time series with an application to daily mean temperature. *Biometrics*, **71**(2), 323–332.

Erhardt, T. M., Czado, C., and Schepsmeier, U. (2015b). Spatial composite likelihood inference using local C-vines. *Journal of Multivariate Analysis*, **138**, 74–88.

Favre, A. C., Adlouni, S. E., Perreault, L., Thiémonge, N., and Bobée, B. (2004). Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, **40**(1), 1–12.

Feehley, C. E., Zheng, C., and Molz, F. J. (2000). A dual-domain mass transfer approach for modeling solute transport in heterogeneous aquifers: Application to the macrodispersion experiment (MADE) site. *Water Resources Research*, **36**(9), 2501–2515.

Fiori, A., Janković, I., Dagan, G., and Cvetković, V. (2007). Ergodic transport through aquifers of non-Gaussian log conductivity distribution and occurrence of anomalous behavior. *Water Resources Research*, **43**(9).

Fiori, A., Dagan, G., Jankovic, I., and Zarlenga, A. (2013). The plume spreading in the MADE transport experiment: Could it be predicted by stochastic models? *Water Resources Research*, **49**(5), 2497–2507.

Fiori, A., Zarlenga, A., Jankovic, I., and Dagan, G. (2017). Solute transport in aquifers: The comeback of the advection dispersion equation and the First Order Approximation. *Advances in Water Resources*, **110**, 349–359.

Fiori, A., Zarlenga, A., Bellin, A., Cvetkovic, V., and Dagan, G. (2019). Groundwater contaminant transport: Prediction under uncertainty, with application to the MADE transport experiment. *Frontiers in Environmental Science*, **7**(JUN).

Fogg, G. E. (1996). Transition Probability-Based Indicator Geostatistics. *Mathematical Geology*, **28**(4), 453–476.

Fogg, G. E., Noyes, C. D., and Carle, S. F. (1998). Geologically based model of heterogeneous hydraulic conductivity in an alluvial setting. *Hydrogeology Journal*, **6**(1), 131–143.

Freeze, R. A. (1975). A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media. *Water Resources Research*, **11**(5), 725–741.

Freyberg, D. L. (1986). A natural gradient experiment on solute transport in a sand aquifer: 2. Spatial moments and the advection and dispersion of nonreactive tracers. *Water Resources Research*, **22**(13), 2031–2046.

Gelhar, L. W. (1977). Effects of hydraulic conductivity variations on groundwater flows. In *Proceeding Second International Symposium on Stochastic HydraulicsAt*, pages 409–431.

Gelhar, L. W. (1986). Stochastic subsurface hydrology from theory to applications. *Water Resources Research*, **22**(9 S), 135S–145S.

Gelhar, L. W. and Axness, C. L. (1983). Three-dimensional stochastic analysis of macrodispersion in aquifers. *Water Resources Research*, **19**(1), 161–180.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**(6), 721–741.

Gilliom, R. J. and Helsel, D. R. (1986). Estimation of Distributional Parameters for Censored Trace Level Water Quality Data: 1. Estimation Techniques. *Water Resources Research*, **22**(2), 135–146.

Giorgi, E., Diggle, P. J., Snow, R. W., and Noor, A. M. (2018). Geostatistical Methods for Disease Mapping and Visualisation Using Data from Spatio-temporally Referenced Prevalence Surveys. *International Statistical Review*, **86**(3), 571–597.

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

Gómez-Hernández, J. J., Journel, A. G., Gomez-Hernandez, J. J., and Journel, A. G. (1993). Joint sequential simulation of multiGaussian fields. In A. Soares, editor, *Geostatistics Troia '92. Vol. 1*, pages 85–94. Springer, Dordrecht.

Gómez-Hernández, J. J., Butler, J. J., Fiori, A., Bolster, D., Cvetkovic, V., Dagan, G., and Hyndman, D. (2017). Introduction to special section on Modeling highly heterogeneous aquifers: Lessons learned in the last 30 years from the MADE experiments and others. *Water Resources Research*, **53**(4), 2581–2584.

Gong, R., Haslauer, C. P., Chen, Y., and Luo, J. (2013). Analytical relationship between Gaussian and transformed-Gaussian spatially distributed fields. *Water Resources Research*, **49**(3), 1735–1740.

Graham, W. and McLaughlin, D. (1989). Stochastic analysis of nonstationary subsurface solute transport: 2. Conditional moments. *Water Resources Research*, **25**(11), 2331–2355.

Gräler, B. (2014). Modelling skewed spatial random fields through the spatial vine copula. *Spatial Statistics*, **10**, 87–102.

Guan, J., Molz, F. J., Zhou, Q., Liu, H. H., and Zheng, C. (2008). Behavior of the mass transfer coefficient during the MADE-2 experiment: New insights. *Water Resources Research*, **44**(2).

Guthke, P. (2013). *Non-multi Gaussian Spatial Structures: Process Driven Natural Genesis, Manifestation, Modeling Approaches, and Influences on Dependent Processes*, volume Heft 220. Institut für Wasser- und Umweltsystemmodellierung, Stuttgart, mitteilung edition.

Guthke, P. and Bárdossy, A. (2017). On the link between natural emergence and manifestation of a fundamental non-Gaussian geostatistical property: Asymmetry. *Spatial Statistics*, **20**, 1–29.

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

Hansen, T. M., Vu, L. T., Mosegaard, K., and Cordua, K. S. (2018). Multiple point statistical simulation using uncertain (soft) conditional data. *Computers and Geosciences*, **114**, 1–10.

Harvey, C. and Gorelick, S. M. (2000). Rate-limited mass transfer or macrodispersion: Which dominates plume evolution at the Macrodispersion Experiment (MADE) site? *Water Resources Research*, **36**(3), 637–650.

Haslauer, C. (2011). *Analysis of Real-World Spatial Dependence of Subsurface Hydraulic Properties Using Copulas With a Focus on Solute Transport Behaviour*. Institut für Wasserbau, Stuttgart, online edition.

Haslauer, C. P., Guthke, P., Brdossy, A., and Sudicky, E. A. (2012). Effects of non-Gaussian copula-based hydraulic conductivity fields on macrodispersion. *Water Resources Research*, **48**(7).

Haslauer, C. P., Bárdossy, A., and Sudicky, E. A. (2017a). Detecting and modelling structures on the micro and the macro scales: Assessing their effects on solute transport behaviour. *Advances in Water Resources*, **107**, 439–450.

Haslauer, C. P., Meyer, J. R., Bárdossy, A., and Parker, B. L. (2017b). Estimating a Representative Value and Proportion of True Zeros for Censored Analytical Data with Applications to Contaminated Site Assessment. *Environmental Science and Technology*, **51**(13), 7502–7510.

Hörning, S. and Bárdossy, A. (2018). Phase annealing for the conditional simulation of spatial random fields. *Computers and Geosciences*, **112**, 101–111.

Jahn, J., Klose, J., and Merkel, A. (1992). On the Application of a Method of Reference Point Approximation to Bicriterial Optimization Problems in Chemical Engineering. In *Advances in Optimization*, pages 478–491. Springer, Berlin, Heidelberg.

Jankovic, I., Maghrebi, M., Fiori, A., and Dagan, G. (2017). When good statistical models of aquifer heterogeneity go right: The impact of aquifer permeability structures on 3D flow and transport. *Advances in Water Resources*, **100**, 199–211.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall/CRC.

Joe, H. (2014). *Dependence modeling with copulas*. CRC Press.

Journel, A. G. (1989). *Fundamentals of Geostatistics in Five Lessons*. American Geophysical Union.

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

Julian, H. E., Boggs, J. M., Zheng, C., and Feehley, C. E. (2001). Numerical simulation of a natural gradient tracer experiment for the natural attenuation study: Flow and physical transport. *Ground Water*, **39**(4), 534–545.

Kazianka, H. and Pilz, J. (2010). Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment*, **24**(5), 661–673.

Kazianka, H. and Pilz, J. (2011). Bayesian spatial modeling and interpolation using copulas. *Computers and Geosciences*, **37**(3), 310–319.

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

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, **220**(4598).

Kitanidis, P. K. (1988). Prediction by the method of moments of transport in a heterogeneous formation. *Journal of Hydrology*, **102**(1-4), 453–473.

Laloy, E., Hérault, R., Jacques, D., and Linde, N. (2018). Training-Image Based Geostatistical Inversion Using a Spatial Generative Adversarial Neural Network. *Water Resources Research*, **54**(1), 381–406.

Lauzon, D. and Marcotte, D. (2019). Calibration of random fields by FFTMA-SA. *Computers and Geosciences*, **127**, 99–110.

Le Borgne, T., Dentz, M., and Carrera, J. (2008). Spatial Markov processes for modeling Lagrangian particle dynamics in heterogeneous porous media. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **78**(2).

LeBlanc, D. R., Garabedian, S. P., Hess, K. M., Gelhar, L. W., Quadri, R. D., Stollenwerk, K. G., and Wood, W. W. (1991). Large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachusetts: 1. Experimental design and observed tracer movement. *Water Resources Research*, **27**(5), 895–910.

Li, J. (2010). *Jing Li Application of Copulas as a New Geostatistical Tool*. Institut für Wasserbau, Stuttgart, online edition.

Li, L., Zhou, H., and Gómez-Hernández, J. J. (2011). A comparative study of three-dimensional hydraulic conductivity upscaling at the macro-dispersion experiment (MADE) site, Columbus Air Force Base, Mississippi (USA). *Journal of Hydrology*, **404**(3-4), 278–293.

Linde, N., Lochbühler, T., Dogan, M., and Van Dam, R. L. (2015). Tomogram-based comparison of geostatistical models: Application to the Macrodispersion Experiment (MADE) site. *Journal of Hydrology*, **531**, 543–556.

Liu, G., Chen, Y., and Zhang, D. (2008). Investigation of flow and transport processes at the MADE site using ensemble Kalman filter. *Advances in Water Resources*, **31**(7), 975–986.

Liu, G., Butler, J. J., Bohling, G. C., Reboulet, E., Knobbe, S., and Hyndman, D. W. (2009). A new method for high-resolution characterization of hydraulic conductivity. *Water Resources Research*, **45**(8).

Liu, G., Zheng, C., Tick, G. R., Butler, J. J., and Gorelick, S. M. (2010). Relative importance of dispersion and rate-limited mass transfer in highly heterogeneous porous media: Analysis of a new tracer test at the Macrodispersion Experiment (MADE) site. *Water Resources Research*, **46**(3).

Liu, G., Butler, J. J., Reboulet, E., and Knobbe, S. (2012). Bestimmung von Vertikalprofilen der hydraulischen Durchlässigkeit mit Direct Push-Methoden. *Grundwasser*, **17**(1), 19–29.

Liu, S., Lu, J. C., Kolpin, D. W., and Meeker, W. Q. (1997). Analysis of environmental data with censored observations. *Environmental Science and Technology*, **31**(12), 3358–3362.

Llopis-Albert, C. and Capilla, J. E. (2009). Gradual conditioning of non-Gaussian transmissivity fields to flow and mass transport data: 3. Application to the Macrodispersion Experiment (MADE-2) site, on Columbus Air Force Base in Mississippi (USA). *Journal of Hydrology*, **371**(1-4), 75–84.

Mackay, D. M., Freyberg, D. L., Roberts, P. V., and Cherry, J. A. (1986). A natural gradient experiment on solute transport in a sand aquifer: 1. Approach and overview of plume movement. *Water Resources Research*, **22**(13), 2017–2029.

Marcotte, D. (1996). Fast variogram computation with FFT. *Computers and Geosciences*, **22**(10), 1175–1186.

Marcotte, D. and Gloaguen, E. (2008). a Class of Spatial Multivariate Models Based on Copulas. In *Proceedings of the eighth international geostatistics congress*, pages 177–186. Geostats.

Mariethoz, G. and Caers, J. (2014). *Multiple-point Geostatistics: Stochastic Modeling with Training Images*. John Wiley & Sons.

Mariethoz, G., Renard, P., and Straubhaar, J. (2010). The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, **46**(11).

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, **5**(3), 439–468.

Meerschaert, M. M., Dogan, M., Van Dam, R. L., Hyndman, D. W., and Benson, D. A. (2013). Hydraulic conductivity fields: Gaussian or not? *Water Resources Research*, **49**(8), 4730–4737.

Mejía, J. M. and Rodríguez-Iturbe, I. (1974). On the synthesis of random field sampling from the spectrum: An application to the generation of hydrologic spatial processes. *Water Resources Research*, **10**(4), 705–711.

Nelsen, R. B. (2000). *An Introduction to Copulas*, volume 95. Springer Publishing Company, Incorporated.

Nur, D., Mengersen, K. L., and Wolff, R. C. (2005). Phase randomization: A convergence diagnostic test for MCMC. *Australian and New Zealand Journal of Statistics*, **47**(3), 309–323.

Pirot, G., Renard, P., Huber, E., Straubhaar, J., and Huggenberger, P. (2015). Influence of conceptual model uncertainty on contaminant transport forecasting in braided river aquifers. *Journal of Hydrology*, **531**, 124–141.

Pollock, D. W. (1988). Semianalytical Computation of Path Lines for Finite-Difference Models. *Groundwater*, **26**(6), 743–750.

Rajaram, H. (2016). Debates—Stochastic subsurface hydrology from theory to practice: Introduction. *Water Resources Research*, **52**(12), 9215–9217.

Rehfeldt, K. R., Boggs, J. M., and Gelhar, L. W. (1992). Field study of dispersion in a heterogeneous aquifer: 3. Geostatistical analysis of hydraulic conductivity. *Water Resources Research*, **28**(12), 3309–3324.

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

Ronayne, M. J., Gorelick, S. M., and Zheng, C. (2010). Geological modeling of submeter scale heterogeneity and its influence on tracer transport in a fluvial aquifer. *Water Resources Research*, **46**(10).

Rubin, Y. (1990). Stochastic modeling of macrodispersion in heterogeneous porous media. *Water Resources Research*, **26**(1), 133–141.

Salamon, P., Fernàndez-Garcia, D., and Gómez-Hernández, J. J. (2007). Modeling tracer transport at the MADE site: The importance of heterogeneity. *Water Resources Research*, **43**(8).

Salvadori, G. and De Michele, C. (2004). Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. *Water Resources Research*, **40**(12), 1–17.

Scheidegger, A. E. (1961). General theory of dispersion in porous media. *Journal of Geophysical Research*, **66**(10), 3273–3278.

Schniger, A., Nowak, W., and Hendricks Franssen, H. J. (2012). Parameter estimation by ensemble Kalman filters with transformed data: Approach and application to hydraulic tomography. *Water Resources Research*, **48**(4).

Schumer, R., Benson, D. A., Meerschaert, M. M., and Baeumer, B. (2003). Fractal mobile/immobile solute transport. *Water Resources Research*, **39**(10).

Sklar, A. (1959). *Fonctions de Répartition àn Dimensions et Leurs Marges*, volume 8. Université Paris 8.

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, **34**(1), 1–21.

Sudicky, E. A. (1986). A natural gradient experiment on solute transport in a sand aquifer: Spatial variability of hydraulic conductivity and its role in the dispersion process. *Water Resources Research*, **22**(13), 2069–2082.

Sudicky, E. A., Illman, W. A., Goltz, I. K., Adams, J. J., and McLaren, R. G. (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**(1).

Sugimoto, T., Bárdossy, A., Pegram, G. G., and Cullmann, J. (2016). Investigation of hydrological time series using copulas for detecting catchment characteristics and anthropogenic impacts. *Hydrology and Earth System Sciences*, **20**(7), 2705–2720.

Tahmasebi, P. and Sahimi, M. (2016a). Enhancing multiple-point geostatistical modeling: 1. Graph theory and pattern adjustment. *Water Resources Research*, **52**(3), 2074–2098.

Tahmasebi, P. and Sahimi, M. (2016b). Enhancing multiple-point geostatistical modeling: 2. Iterative simulation and multiple distance function. *Water Resources Research*, **52**(3), 2099–2122.

Tompson, A. F. and Gelhar, L. W. (1990). Numerical simulation of solute transport in three-dimensional, randomly heterogeneous porous media. *Water Resources Research*, **26**(10), 2541–2562.

Webster, R. and Oliver, M. A. (2008). *Geostatistics for Environmental Scientists: Second Edition*. John Wiley & Sons.

Xiao, B., Haslauer, C., and Bohling, G. (2019). Comparison of multivariate spatial dependence structures of DPIL and flowmeter hydraulic conductivity data sets at the MADE site. *Water (Switzerland)*, **11**(7), 1420.

Ye, J., Lazar, N. A., and Li, Y. (2011). Sparse geostatistical analysis in clustering fMRI time series. *Journal of Neuroscience Methods*, **199**(2), 336–345.

Zhang, C., Song, X., and Azevedo, L. (2021). U-net generative adversarial network for subsurface facies modeling. *Computational Geosciences*, **25**(1), 553–573.

Zhang, D. and Neuman, S. P. (1995). Eulerian-Lagrangian Analysis of Transport Conditioned on Hydraulic Data: 1. Analytical-Numerical Approach. *Water Resources Research*, **31**(1), 39–51.

Zhang, Y. and Benson, D. A. (2008). Lagrangian simulation of multidimensional anomalous transport at the MADE site. *Geophysical Research Letters*, **35**(7).

Zheng, C., Bianchi, M., and Gorelick, S. M. (2011). Lessons learned from 25 years of research at the MADE site. *Ground Water*, **49**(5), 649–662.

Zhou, H., Li, L., Franssen, H. J. H., and Gómez-Hernández, J. J. (2012). Pattern Recognition in a Bimodal Aquifer Using the Normal-Score Ensemble Kalman Filter. *Mathematical Geosciences*, **44**(2), 169–185.

Zinn, B. and Harvey, C. F. (2003). When good statistical models of aquifer heterogeneity go bad: A comparison of flow, dispersion, and mass transfer in connected and multivariate Gaussian hydraulic conductivity fields. *Water Resources Research*, **39**(3), 1–19.